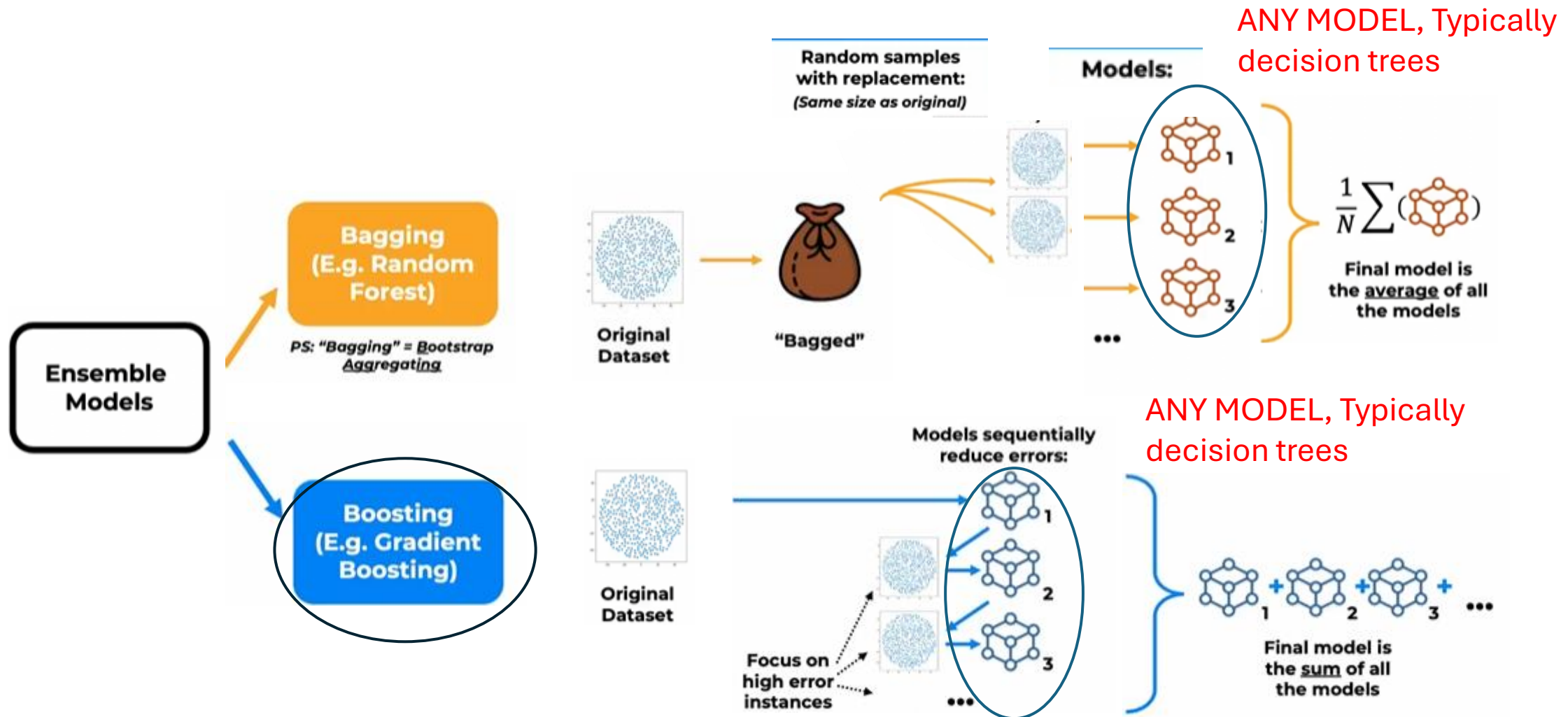


Machine Learning

Ensemble Models



LightGBM (Microsoft, 2016)

Welcome to LightGBM's documentation! 🔗

LightGBM is a gradient boosting framework that uses tree based learning algorithms. It is designed to be distributed and efficient with the following advantages:

- Faster training speed and higher efficiency.
- Lower memory usage.
- Better accuracy.
- Support of parallel, distributed, and GPU learning.
- Capable of handling large-scale data.

<https://lightgbm.readthedocs.io/en/stable/>



LightGBM

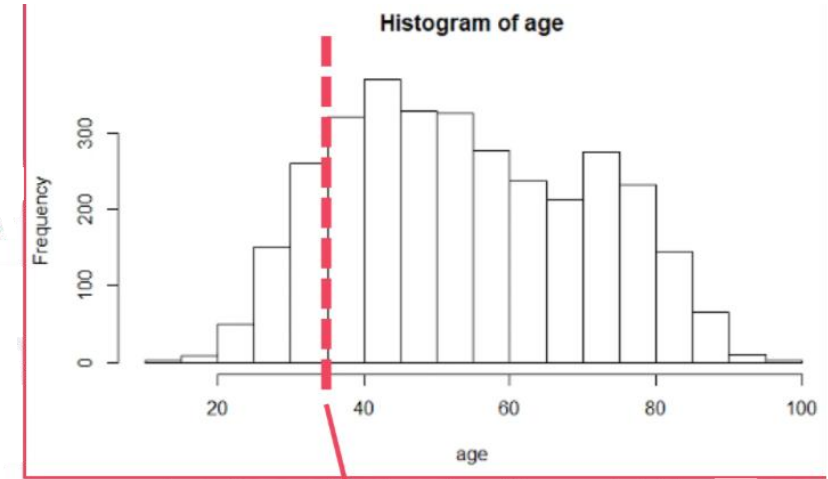
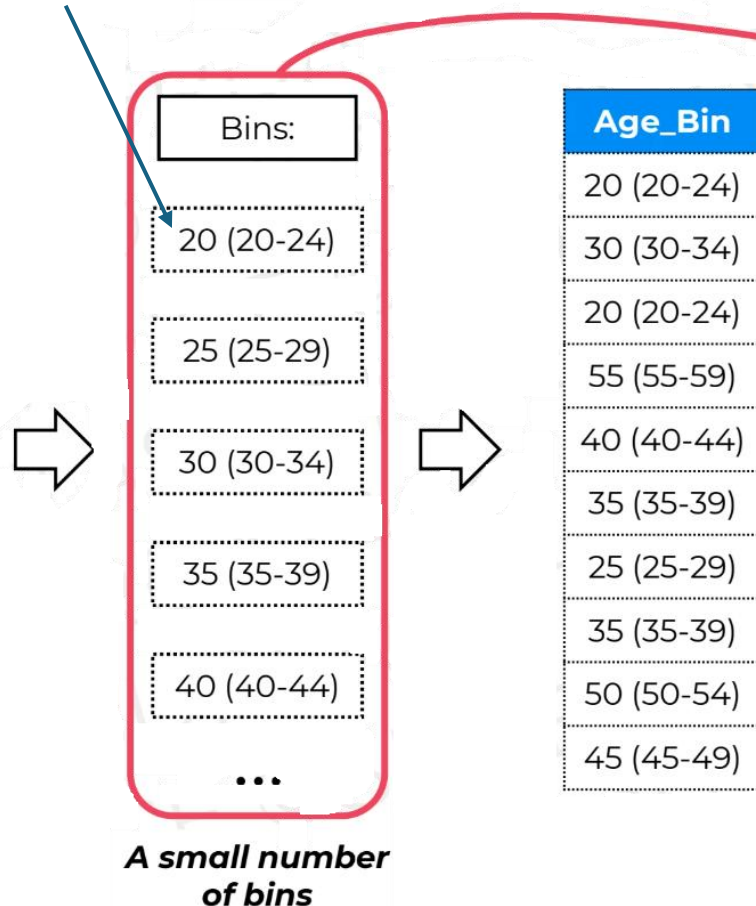
LightGBM compared to XGboost

LightGBM – Histogram Based Training

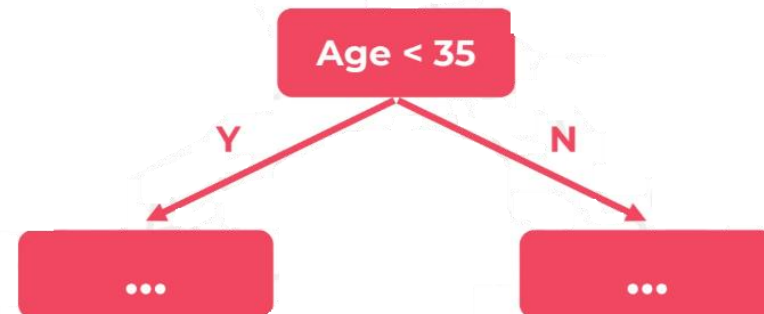
ייצוג ה-bin

Age
22
31
24
56
42
36
28
37
51
49

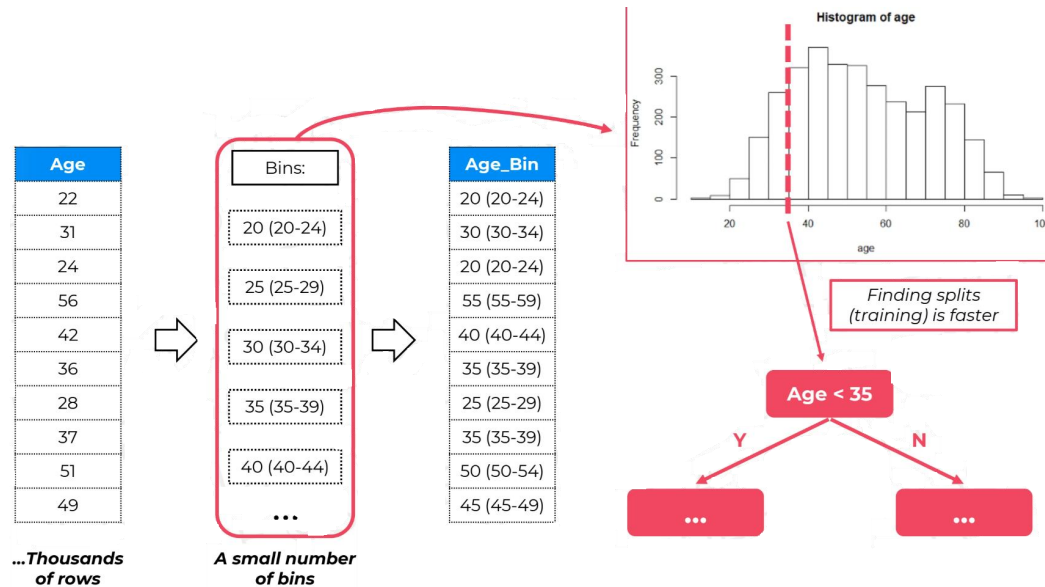
...Thousands of rows



Finding splits (training) is faster



LightGBM – Histogram Based Training



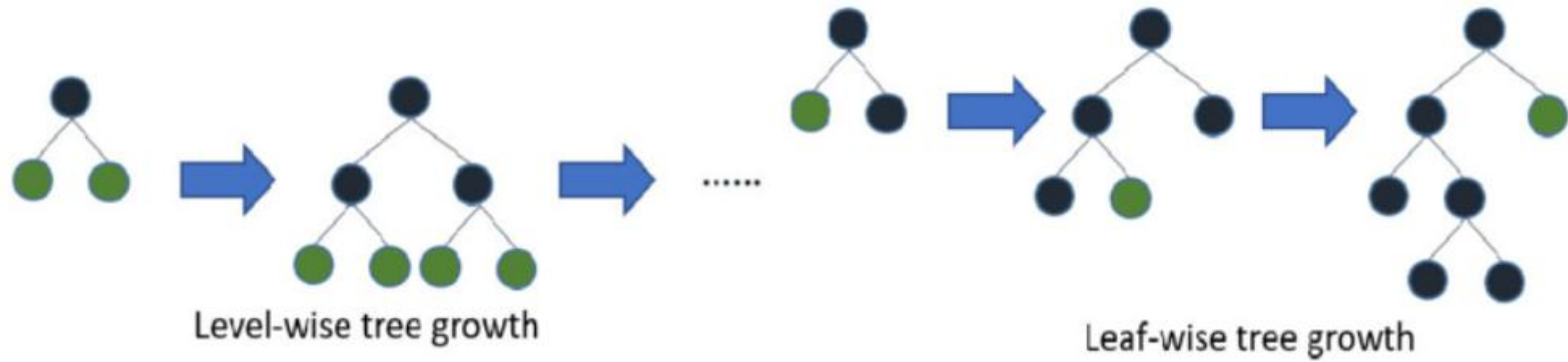
(LightGBM) Histogram-Based Training 🔍

- במקום לבדוק כל ערך רציף אפשרי כדי למצוא את הפיצול הכי טוב — LightGBM מקבץ את הערכים ל-bins (טווחים).
- הוא בונה היסטוגרמה של הערכים, ואז מחפש את נקודת הפיצול הכי טובה בתוך הבינים.

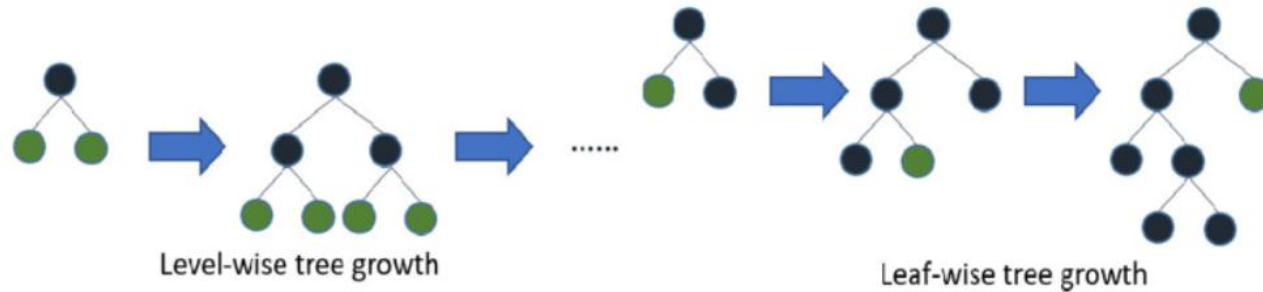
יתרונות: ✓

- הרבה יותר מהיר
- צורך פחות זיכרון
- סקלבי לידאטה עצום

LightGBM – Leaf Wise Tree Growth





LightGBM – Leaf Wise Tree Growth




1. Level-wise (כמו ב־XGBoost ברירת מחדל)

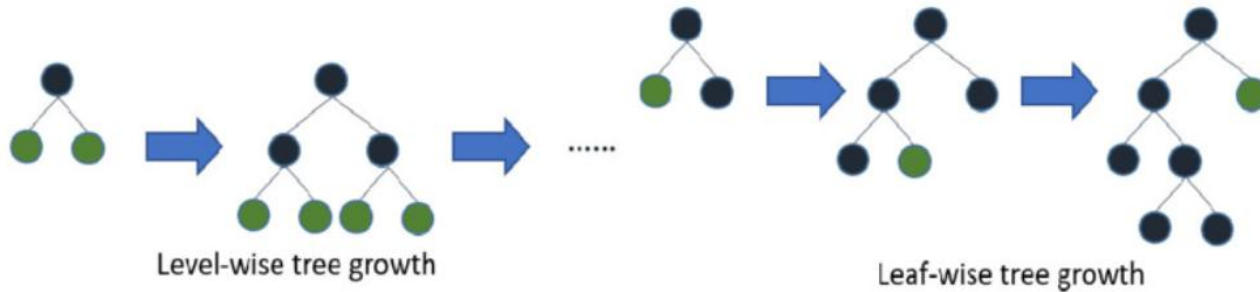
- בונים את העץ **שכבה אחר שכבה** – כלומר כל הרמות נבנות יחד:
- קודם מפצלים את השורש (root)
- אחר כך את כל הצמתים ברמה השנייה
- ואז את כל הרמה השלישית, וכן הלאה...

נבנה בצורה סימטרית יחסית 

יציב 

יכול להיות פחות מדויק, כי לא מתמקד היכן שיש הכי הרבה טעות. 

LightGBM – Leaf Wise Tree Growth



2. Leaf-wise (שיטת LightGBM)

- בכל שלב בוחרים את העלה (leaf) שיביא את שיפור הטעות הכי גדול — ומפצלים רק אותו.
- התוצאה:
- יש ענפים מסוימים בעץ שגדלו עמוק מאוד
- ואחרים יישארו שטוחים

✅ מדויק מאוד – מתמקד באזורים שקשה לחזות
⚠️ נוסה לאוברפיטינג אם לא מגבילים (כי בונה עצים עמוקים)

lightgbm_classifier



CatBoost (Yandex, 2017)

CatBoost – Gradient Boosting חכם לפיצ'רים קטגוריאליים 🌟

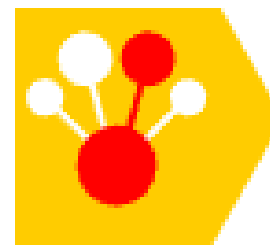
מה זה CatBoost?

- ספריית Boosting שפותחה על ידי Yandex.
- מיועדת בעיקר לדאטה טבלאי (structured/tabular).
- תומכת מצוין בפיצ'רים קטגוריאליים – בלי צורך ב־One-Hot או Label Encoding.

```
from catboost import CatBoostRegressor  
model = CatBoostRegressor(cat_features=[0, 2])  
model.fit(X_train, y_train)
```

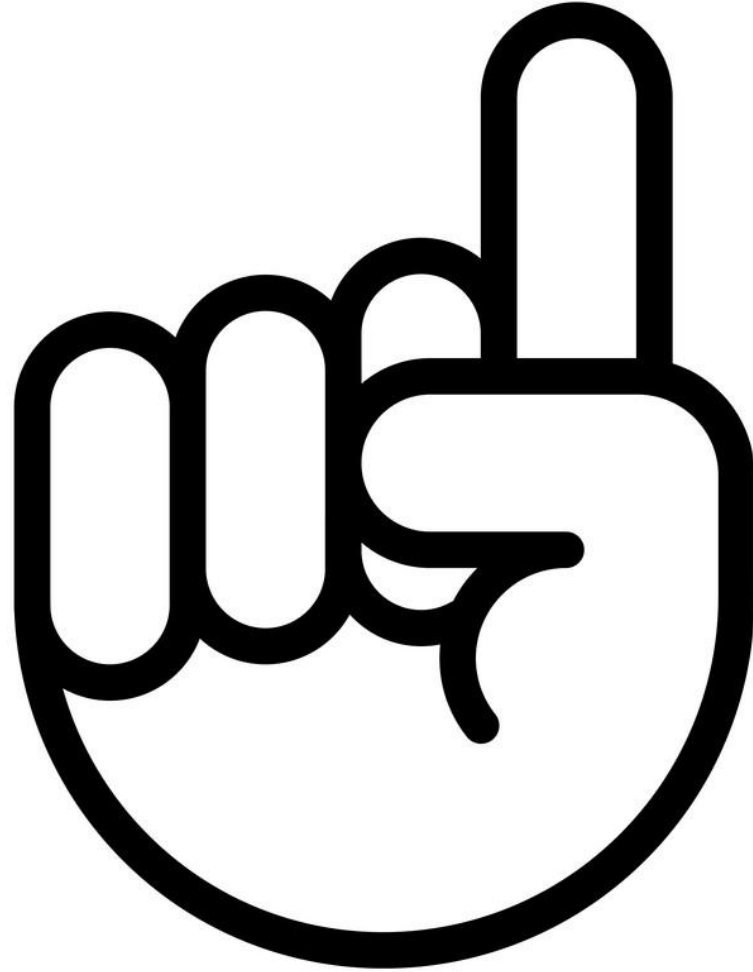
•

<https://catboost.ai/>



CatBoost

catboost_classifier



CatBoost (Yandex, 2017)

```
from catboost import CatBoostRegressor
model = CatBoostRegressor(cat_features=[0, 2])
model.fit(X_train, y_train)
```

"העמודות מספר 0 ו-2 הן קטגוריאליות, ו-CatBoost צריך לטפל בהן בהתאם."

המספרים כאן הם אינדקסים של עמודות, כלומר לפי הסדר ב-`X_train`.
לדוגמה:

Type	Column Name	Index
קטגוריאלי (male, female)	gender	0
מספרי	age	1
קטגוריאלי (north, south)	region	2

אז במקרה הזה `cat_features=[0, 2]` נכון, כי:

- `gender` ו-`region` הן קטגוריות
- `age` הוא רציף (ולא נכנס לרשימה)

Xgboost vs LightGBM vs CatBoost

CatBoost	LightGBM	XGBoost	קריטריון
מהיר מאוד (ב־CPU)	מהיר מאוד (histogram)	מהיר	מהירות 🚀
Symmetric trees	Leaf-wise	Level-wise	בניית עצים 🏗️
כן (מובנה + מתקדם) ✅	כן (מובנה) ✅	לא (דורש encoding) ❌	תמיכה בפיצ'רים קטגוריים 📊
כן ✅ מלא	חלקי	חלקי	טיפול ב־Missing Values 🗣️
נמוכה (Ordered Boosting)	גבוהה (עצים עמוקים)	בינונית	נטייה לאוברפיטינג 📈