CatBoost is an open-source machine learning algorithm for gradient boosting on decision trees, primarily used for various supervised learning tasks, especially with data containing a high number of categorical features

Primary Uses and Applications

CatBoost is a versatile tool applicable across diverse industries for a variety of tasks:

Classification: Predicting which category an observation belongs to, such as in fraud detection or [email spam detection].

Regression: Predicting continuous numerical values, such as in house price prediction, stock market forecasting, or weather prediction.

Ranking: Determining the relative order of items, a key task in search engines (e.g., Yandex search) and recommendation systems.

Other domains:
Recommendation systems for e-commerce or media platforms.
Self-driving cars and personal assistants.
Medical diagnoses by analyzing patient data.
Bot detection (used by Cloudflare).

Key Feature: Native Handling of Categorical Data

The most significant advantage and primary use case for CatBoost is its robust, automatic handling of categorical features (hence "Cat" in the name).

No Preprocessing Needed: Unlike other popular algorithms like XGBoost, which require manual encoding (e.g., one-hot encoding or label encoding), CatBoost can use non-numeric data types directly.
Reduces Overfitting and Data Leakage: It uses an innovative technique called Ordered Boosting with random permutations to transform categorical variables into numerical ones during training, preventing target leakage and the prediction shift that can occur with traditional methods.

Improved Accuracy: This native handling often leads to better performance and more robust models "out of the box" with minimal hyperparameter tuning.
In essence, data scientists use CatBoost to save significant time on data preprocessing, reduce the risk of errors, and achieve high-performance results, especially when dealing with the heterogeneous, tabular data common in real-world business problems.

XGBoost-

Traditional machine learning models like decision trees and random forests are easy to interpret but often struggle with accuracy on complex datasets.

XGBoost short form for eXtreme Gradient Boosting is an advanced machine learning algorithm designed for efficiency, speed and high performance

It is an optimized implementation of Gradient Boosting and is a type of ensemble learning method that combines multiple weak models to form a stronger model.

XGBoost uses decision trees as its base learners and combines them sequentially to improve the model's performance.

Each new tree is trained to correct the errors made by the previous tree and this process is called boosting.
It has built-in parallel processing to train models on large datasets quickly.
XGBoost also supports customizations allowing users to adjust model parameters to optimize performance based on the specific problem.

How XGBoost Works?

It builds decision trees sequentially with each tree attempting to correct the mistakes made by the previous one. The process can be broken down as follows:

Start with a base learner: The first model decision tree is trained on the data. In regression tasks this base model simply predicts the average of the target variable.

Calculate the errors:
After training the first tree the errors between the predicted and actual values are calculated.

Train the next tree:
The next tree is trained on the errors of the previous tree.
This step attempts to correct the errors made by the first tree.

Repeat the process: This process continues with each new tree trying to correct the errors of the previous trees until a stopping criterion is met.

Combine the predictions: The final prediction is the sum of the predictions from all the trees.