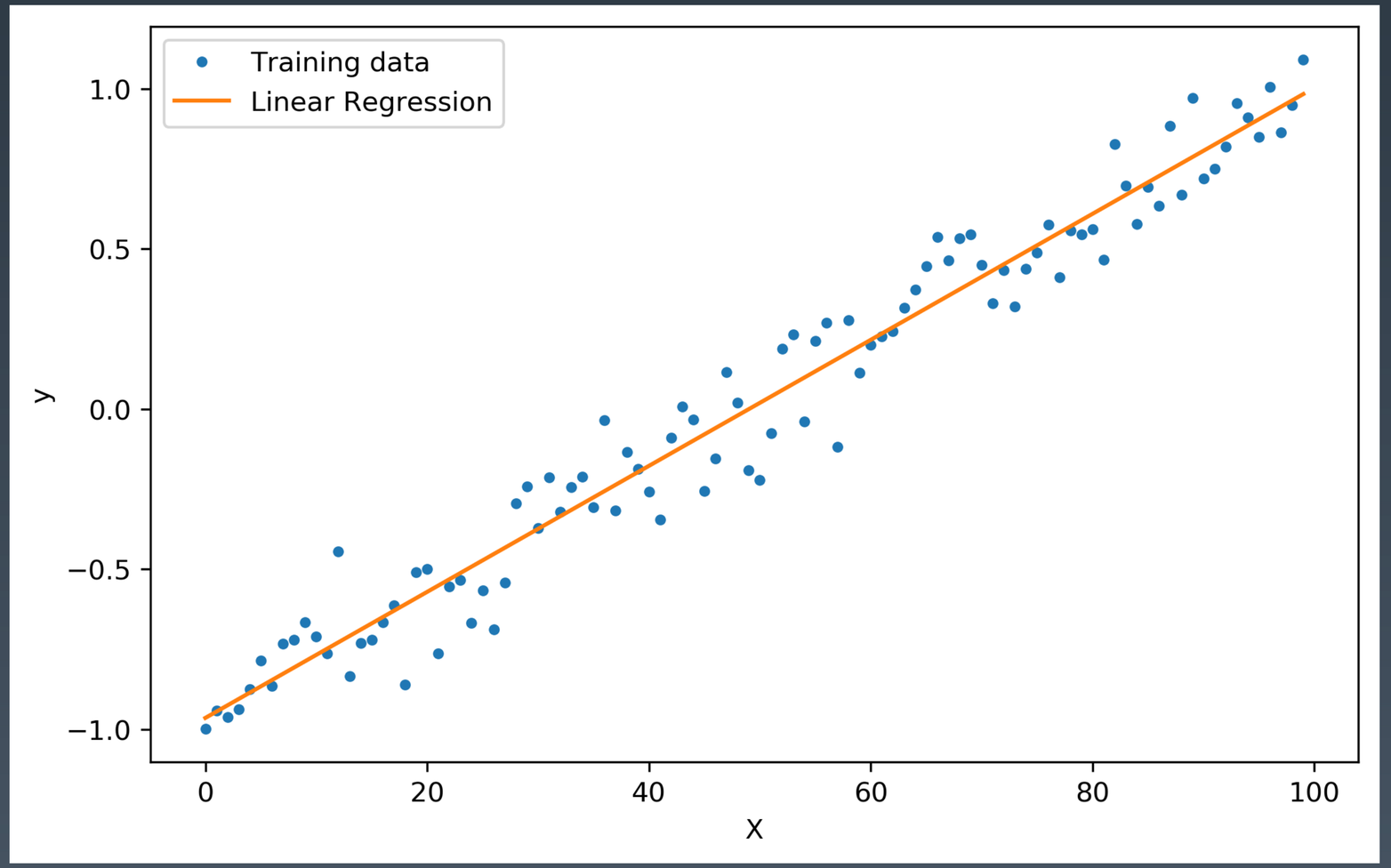
极客大学

机器学习小课堂之
初探 LightGBM

王然/众微科技 AI Lab 负责人

树模型和线性模型：线性模型

极客大学



树模型和线性模型：线性模型

极客大学

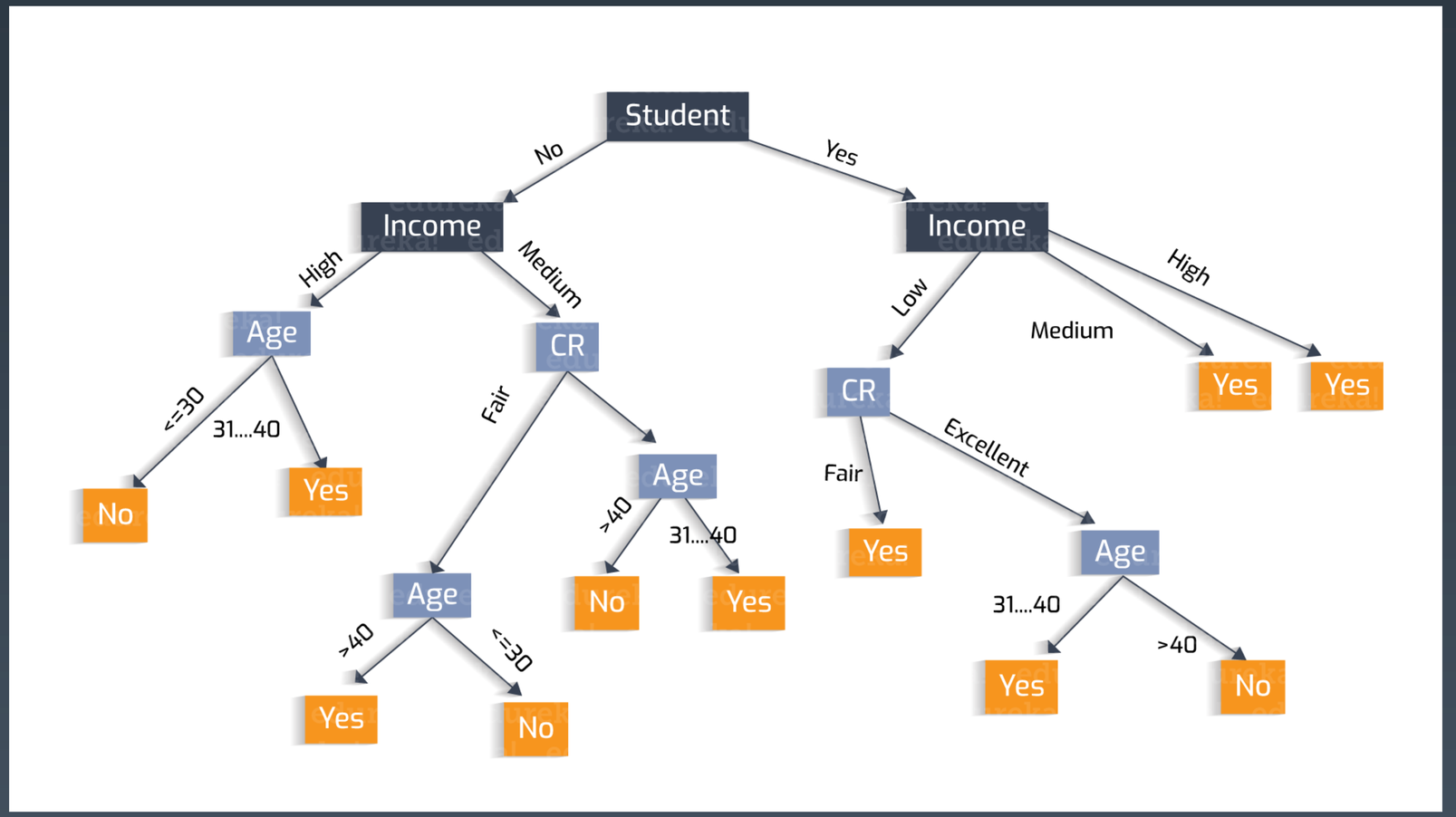
1.非线性关系

2.交叉效应

3.计算复杂度

4.变量选择问题

树模型和线性模型：决策树模型



树模型和线性模型：决策树模型的优点与问题

极客大学

优点

- 可以捕捉非线性关系（？）
- 可以捕捉交叉效应
- 训练效率高
- 自带变量选择（？）

问题

- 准确率很低
- 模型形式过于简单（？）
- 不稳定

集成树模型：随机森林

极客大学

1. 基本想法：如果一棵树不行，那么很多树是否可以呢?

2.主要流程：随机抽取部分观测和变量，拟合树模型，采用投票方式决定结果

3.主要变种：ExtraTrees

4.问题：每一棵树和前一棵树没有关系

集成树模型：GBDT

极客大学

Algorithm 1: Gradient_Boost

1 $F_0(\mathbf{x}) = \arg\min_\rho \sum_{i=1}^{N} L(y_i, \rho)$

2 For $m = 1$ to M do:

3 $\tilde{y}_i = -\left[\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, \; i = 1, N$

4 $\mathbf{a}_m = \arg\min_{\mathbf{a},\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(\mathbf{x}_i; \mathbf{a})]^2$

5 $\rho_m = \arg\min_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + \rho h(\mathbf{x}_i; \mathbf{a}_m))$

6 $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}; \mathbf{a}_m)$

7 endFor

end Algorithm

集成树模型：LightGBM

极客大学

Algorithm 2: Gradient-based One-Side Sampling

Input: I: training data, d: iterations
Input: a: sampling ratio of large gradient data
Input: b: sampling ratio of small gradient data
Input: $loss$: loss function, L: weak learner
models $\leftarrow$ {}, fact $\leftarrow \frac{1-a}{b}$
topN $\leftarrow$ a $\times$ len(I) , randN $\leftarrow$ b $\times$ len(I)
for $i = 1$ ***to*** d **do**
    preds $\leftarrow$ models.predict(I)
    g $\leftarrow$ $loss$(I, preds), w $\leftarrow$ {1,1,...}
    sorted $\leftarrow$ GetSortedIndices(abs(g))
    topSet $\leftarrow$ sorted[1:topN]
    randSet $\leftarrow$ RandomPick(sorted[topN:len(I)], randN)
    usedSet $\leftarrow$ topSet + randSet
    w[randSet] $\times =$ fact $\triangleright$ Assign weight $fact$ to the small gradient data.
    newModel $\leftarrow$ L(I[usedSet], $-$ g[usedSet], w[usedSet])
    models.append(newModel)

LightGBM 实战和调参

极客大学

1. 跑通 Demo

2. 重要参数和注意事项

3. 调参方法论

THANKS!