Євчик Олексій, лаб 7.
Контрольні питання — відповіді

1. Навіщо в Random Forest одночасно використовують бутстреп за об’єктами та випадковий вибір ознак?

Тому що комбінування цих двох джерел випадковості:

- зменшує кореляцію між деревами,
- збільшує різноманітність моделей,
- зменшує дисперсію загального ансамблю,

що робить Random Forest стійкішим, менш схильним до переобучення та загалом більш точним.

---

2. Чим OOB-оцінка корисна на практиці?

OOB-оцінка:

- дозволяє оцінити якість моделі без виділення окремого валідаційного набору;
- майже безкоштовна, бо використовує ті спостереження, які не потрапили у бутстреп зразок для окремого дерева;
- добре узгоджується з реальною помилкою моделі.

Це робить її ідеальним способом швидко оцінювати гіперпараметри.

3. Як стартово вибирати mtry в класифікації та регресії?

Стандартні рекомендації:

- Класифікація:

$$mtry = \sqrt{p}$$

де p — кількість ознак.

- Регресія:

$$mtry = \frac{p}{3}$$

Це стартові значення, далі їх оптимізують експериментально.

4. Коли застосовувати permutation-важливість замість Gini-важливості?

Permutation importance застосовують, коли:

- є корельовані ознаки (Gini їх переоцінює),
- нам потрібна реальна, а не структурна важливість,
- важлива стабільність і чесність оцінки важливості.

Gini важливість швидша, але спотворюється через кореляції та різні масштаби ознак.

5. Які підходи використовувати при дисбалансі класів?

Основні прийоми:

- Зміна ваг класів (class.weights)
- Пересемплінг:
    - oversampling (SMOTE, ROSE)
    - undersampling
- Поріг ймовірності (threshold tuning)
- Стратегія balanced subsample у Random Forest

- Метрики, нечутливі до дисбалансу: AUC, F1, balanced accuracy

6. Як читати PDP і чому важлива кореляція ознак?

PDP — Partial Dependence Plot показує, як середній прогноз моделі змінюється при зміні однієї або двох ознак.

Але важливо знати:

- якщо ознаки сильно корельовані, то PDP може показувати нереалістичні комбінації значень;
- у таких випадках PDP може бути спотвореним і вводити в оману.

Тому для корельованих ознак краще використовувати ICE-плоти або ALE-плоти.

7. Як змінюється bias/variance при зміні mtry і min.node.size?

mtry

- Менший mtry → дерева більш різні →
  менша дисперсія ансамблю
  більший bias (бо моделі слабші)
- Більший mtry → дерева схожіші →
  менший bias
  більша дисперсія (моделі можуть переобучуватися)

min.node.size

- Малий min.node.size → глибші дерева →
  менший bias
  більша variance

- Великий min.node.size → дрібні дерева обрізаються →
  менша variance
  більший bias