



Data Science Intern: Talablar va vazifalar

Asosiy bilimlar

- pandas
- numpy
- matplotlib
- scikit-learn
- ehtimollik nazariyasi va statistika asoslari

Plus bo'ladi

- SQL bilimiga ega bo'lish
- pytorch bilan tajriba

Asosiy vosita: Python

Topshiriqlar:

1-qism. Analitik fikrlash

Quyida internet-do'kon ma'lumotlari keltirilgan. Har bir satr — alohida foydalanuvchi.

user_id	source	visits	purchases	revenue
1	ads	5	1	120
2	ads	3	0	0
3	organic	4	1	80
4	organic	6	2	300

5	ads	2	0	0
6	organic	1	0	0

Savollar:

1. ads va organic uchun purchase conversion ko'rsatkichini hisoblang.
2. Qaysi manba yaxshiroq ko'rinadi va nega bu xulosa noto'g'ri bo'lishi mumkin?
3. Qo'shimcha ravishda kuzatish kerak bo'lgan yana 2 ta metrikani ayting.

2-qism. Ehtimollik nazariyasi

Bitta tashrifda xarid ehtimoli — 0.2.

3 ta tashrif davomida kamida bitta xarid amalga oshishi ehtimolini hisoblang.

3-qism. Machine Learning (qisqa javob)

1. Data leakage nima va u nima uchun xavfli?
2. train / validation / test nimasi bilan farqlanadi?
3. logistic regression va decision tree o'rtasidagi asosiy farq nimada?
4. Train'da natija yuqori, test'da past bo'lsa — bu nimani anglatadi va nima qilish mumkin?
5. Ma'lumot kam bo'lganda oddiy model murakkab modeldan yaxshiroq ishlashi mumkinligi sababi nima?

4-qism. Python / Amaliy topshiriq

DataFrame df quyidagi ustunlardan iborat:

user_id | age | time_spent | purchased

Vazifalar

- purchased = 0 va purchased = 1 bo'yicha time_spent o'rtacha qiymatini hisoblang
- Xarid qilgan foydalanuvchilar ulushini hisoblang

Izoh bering:

- Qaysi belgilardan (features) foydalanish foydali bo'lishi mumkin
- Qaysi belgilarni ishlatmaslik kerak va sababi