DATA SCIENCE ASOSIDA MA'LUMOTLARNI QAYTA ISHLASH

Mualliflar: **Rahimov Quvvatali Ortiqovich** *t.f. b.f.d. (PhD)*
(FarDU "Axborot texnologiyalari" kafedrasi mudiri).
Yusupov Mirsaid Abdulaziz o'g'li
(FarDU "Amaliy matematika" mutaxassiligi 2 - kurs magistranti).
Umirjonov Lazizjon Akmaljon o'g'li
(FarDU "Amaliy matematika" mutaxassiligi 2 - kurs magistranti).

Kalit so'zlar: *Data Science, ma'lumot, qayta ishlash, filtrlash, himoya qilish.*

Ma'lumotlar metodologiyasining asl tavsifida Fan - loyiha pazandachilik bilan taqqoslanadi va tahlilchi - Bilan oshpaz . Shunga ko'ra, ma'lumotlarni tayyorlash bosqichi mahsulotlarni tayyorlash bilan taqqoslanadi: biz biznes topshirig'ini tahlil qilish bosqichida pishiradigan taomning retsepti haqida qaror qabul qilganimizdan so'ng, biz topishimiz, bir joyda yig'ishimiz, tozalashimiz va tozalashimiz kerak. ingredientlarni kesib oling. Shunga ko'ra, taomning ta'mi ushbu bosqichning qanchalik to'g'ri bajarilganiga bog'liq bo'ladi (deylik, biz retseptni to'g'ri taxmin qildik, ayniqsa jamoat mulkida juda ko'p retseptlar mavjud). Ingredientlar bilan ishlash, ya'ni ma'lumotlarni tayyorlash har doim zargarlik, mashaqqatli va mas'uliyatli ishdir: bitta buzilgan yoki yuvilmagan mahsulot - va barcha ish behuda ketadi.

Ma'lumotlar yig'ish

Kerak bo'lishi mumkin bo'lgan ingredientlar ro'yxatini olganimizdan so'ng , biz muammoni hal qilish uchun ma'lumotlarni qidirishni boshlaymiz va kelajakda ishlashimiz mumkin bo'lgan namunani shakllantiramiz. Bizga namuna nima uchun kerakligini eslang: birinchidan, biz ma'lumotlarni tayyorlash bosqichida ma'lumotlarning tabiati haqida tasavvurga ega bo'lish uchun foydalanamiz, ikkinchidan, biz undan ishlab chiqish va sozlash bosqichlarida sinov va o'quv namunalarini shakllantiramiz. model.

Albatta, biz qattiq NDA keng ko'lamli loyihadagi ma'lumotlar haqida biror narsani tushunishingiz kerak bo'lgan holatlarni qabul qilmaymiz va siz qattiq mijozdan ma'lumotlarni flesh- diskda yoki xatga ilova sifatida. Aytaylik, sizda hamma narsa yaxshi va sizda ma'lumotlarga kirish imkoningiz bor. Ushbu bosqichda quyidagi namunani tayyorlash kerak:

1. umumiy aholining barcha zaruriy xususiyatlarini aks ettirgan
2. ishlash uchun qulay edi, ya'ni juda katta emas edi.

Bu sotsiologiyada, qoida tariqasida, umumiy aholi mavjud emas: biz jamoatchilik fikrini o'rganganimizda, hatto nazariy jihatdan ham hamma odamlardan intervyu olish mumkin emas. Yoki tibbiyotda ma'lum miqdordagi tajriba quyonlari / sichqonlari / chivinlarida yangi dori o'rganilmoqda: tadqiqot guruhidagi har bir qo'shimcha ob'ekt

qimmat, mashaqqatli va qiyin. Biroq, butun aholi biz uchun mavjud bo'lsa ham, katta ma'lumotlar hisoblash uchun tegishli infratuzilmani talab qiladi va uni joylashtirishning o'zi qimmat (biz infratuzilma tayyor va sizning qo'lingizda sozlangan holatlar haqida gapirmayapmiz). Ya'ni, nazariy jihatdan barcha ma'lumotlarni hisoblash mumkin bo'lsa ham, odatda bu uzoq, qimmat va umuman nima uchun ekanligi ma'lum bo'ladi, chunki agar siz kichik bo'lsa ham yuqori sifatli namunani tayyorlasangiz, bularning barchasisiz qila olasiz. masalan, bir necha ming yozuvlardan iborat.

Namuna olishga sarflangan vaqt va kuch ma'lumotlarni o'rganishga ko'proq vaqt sarflashga imkon beradi: masalan, o'zgarib turadigan yoki etishmayotgan ma'lumotlar qimmatli ma'lumotlarni o'z ichiga olishi mumkin, lekin uni millionlab yozuvlar orasidan topish mumkin emas, lekin bir necha minglar orasida umuman topilmaydi .

Ma'lumotlarning reprezentativligini baholash

Tushunish uchun bizga sog'lom fikr va statistika kerak. Kategorik ma'lumotlar uchun biz namunamizda biznes vazifasi uchun muhim bo'lgan har bir xususiyat umumiy populyatsiyadagi kabi bir xil nisbatda ifodalanganligiga ishonch hosil qilishimiz kerak. Misol uchun, agar biz klinikadagi bemorlarning ma'lumotlarini tekshirayotgan bo'lsak va savol barcha yoshdagi odamlarga tegishli bo'lsa, biz faqat bolalarni o'z ichiga olgan namunani sig'dira olmasligimiz mumkin. Tarixiy ma'lumotlar uchun ma'lumotlar o'rganilayotgan xususiyatlar barcha mumkin bo'lgan qiymatlarni oladigan vakillik vaqt oralig'ini qamrab olganligini tekshirishga arziydi. Misol uchun, agar biz davlat organlariga murojaatlarni tahlil qilsak, yanvar oyining birinchi haftasi uchun ma'lumotlar bizga mos kelmaydi , chunki bu vaqtda murojaatlar kamaygan. Raqamli xususiyatlar uchun asosiy statistik ma'lumotlarni hisoblash mantiqan to'g'ri keladi (hech bo'lmaganda bitta nuqta: o'rtacha, median, o'zgaruvchanlik va iloji bo'lsa, umumiy aholining o'xshash statistikasi bilan solishtiring, albatta).

Ma'lumotlarni yig'ishdagi muammolar.

Ko'pincha bizda etarli ma'lumot yo'qligi sodir bo'ladi. Masalan, axborot tizimi o'zgargan va eski tizimdagi ma'lumotlar mavjud emas yoki ma'lumotlar strukturasi boshqacha: yangi kalitlardan foydalaniladi va eski va yangi ma'lumotlar o'rtasidagi munosabatni o'rnatib bo'lmaydi. Ma'lumotlar turli egalar tomonidan saqlanishi va har kim ham uchinchi tomon loyihasi uchun yuklash uchun vaqt va resurslarni sarflashga sozlanishi mumkin bo'lmaganda, tashkiliy muammolar ham odatiy hol emas.

Bunday holatda qanday bo'lish kerak? Ba'zida uning o'rnini topish mumkin bo'ladi: agar yangi pomidor bo'lmasa, konservalar paydo bo'lishi mumkin. Va agar sabzi chirigan bo'lib chiqsa, siz yangi qism uchun bozorga borishingiz kerak. Shunday qilib, bu bosqichda biz oldingi bosqichga qaytishimiz kerak bo'ladi, u erda biz biznes muammosini tahlil qildik va savolni qandaydir tarzda qayta shakllantirish mumkinmi yoki yo'qmi, deb o'ylaymiz: masalan, biz onlayn rejimning qaysi versiyasini aniq aniqlay olmaymiz. do'kon sahifasi mahsulotni yaxshiroq sotadi (aytaylik, savdo ma'lumotlari etarli emas), lekin biz foydalanuvchilarning qaysi sahifaga ko'proq vaqt sarflashini va

qaysi sahifaga kamroq qaytishini (bir necha soniya ichida juda qisqa ko'rish seanslari) aniqlashimiz mumkin .Shundan so'ng siz har bir xususiyatni alohida o'rganishga o'tishingiz mumkin. Asosiy tadqiqot vositasi tavsiflovchi statistika hisoblanadi.

Tekshirish ma'lumotlarini tahlil qilish

Aytaylik, ma'lumotlar qabul qilindi va u umumiy aholini aks ettiradi va belgilangan biznes muammosiga javobni o'z ichiga oladi. Endi ma'lumotlar bizning qo'limizda qanday sifat va ular mo'ljallangan retseptga mos keladimi yoki yo'qligini tushunish uchun ularni tekshirish kerak . Aytaylik, biz allaqachon bir nechta yozuvlar misollarini oldik, kalit nima ekanligini va unda qanday ma'lumotlar turlari mavjud: raqamli, ikkilik, kategorik. Shundan so'ng siz har bir xususiyatni alohida o'rganishga o'tishingiz mumkin. Asosiy tadqiqot vositasi tavsiflovchi statistika hisoblanadi.

Markaziy pozitsiyani baholash.

Tadqiqotning birinchi bosqichida har bir xususiyat uchun qanday qiymatlar xos ekanligini tushunish yaxshi bo'lar edi. Eng oddiy baholash o'rtacha arifmetikdir: oddiy va taniqli ko'rsatkich. Biroq, agar ma'lumotlarning tarqalishi katta bo'lsa, unda o'rtacha ko'rsatkich bizga odatiy qiymatlar haqida ko'p gapirmaydi: masalan, biz kasalxonada ish haqi darajasini tushunishni xohlaymiz. Buning uchun biz hamshiradan bir necha barobar ko'p oladigan barcha xodimlarning, shu jumladan direktorning maoshini qo'shamiz. Olingan arifmetik o'rtacha har qanday xodimlarning (direktordan tashqari) maoshidan yuqori bo'ladi va odatdagi ish haqi haqida bizga hech narsa aytmaydi. Bunday ko'rsatkich faqat ish haqining oshishi haqida g'urur bilan xabar beradigan Sog'liqni saqlash vazirligiga hisobot berish uchun javob beradi. Olingan qiymatga chegara qiymatlari juda ta'sir qiladi. Tartibda _ chet ko'rsatkichlar (atipik, ekstremal qiymatlar) ta'siridan qochish uchun boshqa statistik ma'lumotlardan foydalaniladi: tartiblangan qiymatlarda markaziy qiymat sifatida hisoblangan mediana.

Agar ma'lumotlar ikkilik yoki toifali bo'lsa, qaysi qiymatlar ko'proq va qaysi biri kamroq tarqalganligini bilishga arziydi. Buning uchun rejim ishlatiladi: eng keng tarqalgan qiymat yoki toifa. Bu , boshqa narsalar qatori, namunaning reprezentativligini tushunish uchun foydalidir : masalan, biz bemorlarning tibbiy yozuvlari ma'lumotlarini o'rganib chiqdik va kartalarning ⅔ qismi ayollarga tegishli ekanligini aniqladik. Bu sizni namunani tanlashda xatolik yuz berdimi, degan savol tug'diradi. Kategoriyalarning bir-biriga nisbatan nisbatlarini ko'rsatish uchun ma'lumotlarning grafik tasviri, masalan, chiziqli yoki doiraviy diagrammalar ko'rinishida foydalidir.

MA'LUMOTLARNING O'ZGARUVCHANLIGINI BAHOLASH

Namunamizning odatiy qiymatlari to'g'risida qaror qabul qilganimizdan so'ng, biz atipik qiymatlarni - chetga chiqishni ko'rishimiz mumkin. Chet elliklar bizga ma'lumotlarning sifati haqida biror narsa aytib berishi mumkin: masalan, ular xatolik belgilari bo'lishi mumkin: o'lchamdagi chalkashliklar, o'nlik kasrlarning yo'qolishi yoki

kodlash egri chizig'i. Shuningdek, ular ma'lumotlarning qanchalik o'zgarishi, o'rganilayotgan xususiyatlarning chegaraviy qiymatlari haqida gapirishadi.

Keyinchalik, ma'lumotlar qanchalik farq qilishini umumiy baholashga o'tishingiz mumkin. O'zgaruvchanlik (aka dispersiya) belgi qiymatlari bir-biridan qanchalik farq qilishini ko'rsatadi. O'zgaruvchanlikni o'lchash usullaridan biri belgilarning markaziy qiymatdan odatiy og'ishlarini baholashdir . Bu og'ishlarni o'rtacha hisoblash bizga ko'p narsa bermasligi aniq, chunki salbiy og'ishlar ijobiylarni zararsizlantiradi. O'zgaruvchanlikning eng mashhur o'lchovlari dispersiya va standart og'ish bo'lib , ular og'ishlarning mutlaq qiymatini hisobga oladi (dispersiya - kvadratik og'ishlarning o'rtacha qiymati, standart og'ish - dispersiyaning kvadrat ildizi).

Boshqa yondashuv saralangan ma'lumotlarning tarqalishini hisobga olishga asoslangan (katta ma'lumotlar to'plamlari uchun bu o'lchovlar qo'llanilmaydi, chunki qiymatlar birinchi navbatda saralanishi kerak, bu o'z-o'zidan qimmat) . Misol uchun, foizlar yordamida baholash (siz faqat sentillarni ham topishingiz mumkin). N _ Persentil - ma'lumotlarning kamida N foizi shu qiymat yoki undan kattaroq bo'lgan qiymatdir . Cheklovlarga nisbatan sezgirlikni oldini olish uchun har bir uchidan qiymatlarni olib tashlash mumkin. O'zgaruvchanlikning umumiy o'lchovi 25 va 75 foizlar orasidagi farqdir - kvartillararo diapazon.

ADABIYOTLAR RO‘YHATI:

1. Кукарцев, В. В. Теория баз данных: учебник / В. В. Кукарцев, Р. Ю. Царев, О. А. Антамошкин. — Электрон. текстовые данные. — Красноярск : Сибирский федеральный университет, 2017. — 180 c. — 978-5-7638-3621-9.

2. Киселева Т. В. Программная инженерия. Часть 1 [Elektron resurs] : учебное пособие / Т. В. Киселева. — Ставрополь : Северо-Кавказский федеральный университет, 2017. — 137 с.

3. Analytics Comes of Age. [Elektron resurs]

4. Apache Cassandra. [Elektron resurs]

5. Apache CouchDB. [Elektron resurs]

6. Apache HBase – Apache HBase™ Home. [Elektron resurs]

7. ArangoDB: Multi-model highly available NoSQL database. [Elektron resurs]

8. Azure Cosmos DB. Мультимодельная, глобально распределенная служба базы данных для любого масштаба [Elektron resurs]

9. Biehn Neil. The Missing V’s in Big Data: Viability and Value / Neil Biehn. [Elektron resurs]

10. Big Data Analytics Landscape 2019. [Elektron resurs]

11. Big Data Executive Survey 2017. Executive Summary of Findings. [Elektron resurs]

12. Ortiqovich, Q. R. ., & Ismoiljon o‘g‘li, A. S. . (2023). Learning the Diffusion Equation Using Python. Best Journal of Innovation in Science, Research and Development, 2(5), 173–180. Retrieved from

13. Ortiqovich, Q. R. ., & Ismoiljon o‘g‘li, A. S. . (2023). Learning the Diffusion Equation Using Python. Best Journal of Innovation in Science, Research and Development, 2(5), 173–180. Retrieved from

14. Абдулазиз угли, Ю. М., Каримбердиевич, О. М., & Махамадин угли, Ё. А. (2022). АЛГОРИТМЫ РАСПОЗНОВАНИЯ РЕЧИ И КЛАССИФИКАЦИЯ МЕТОДОВ РАСПОЗНОВАНИЯ РЕЧИ. CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES, 3(10), 15-19. Retrieved from https://cajmtcs.centralasianstudies.org/index.php/CAJMTCS/article/view/240MORE CITATION FORMATS

15. Каримов Ш.Т., Хайдарова С.А. Численное решение периодических уравнений с дробно-интегральным оператором вейля в главной части.//Fars Int J Soc Sci Hum 10(12);2022. Publishing centre of Finland. С.152-157.

16. Фармонов Ш., Хайдарова С. Обобщенный метод Бубнова-Галеркина для уравнений с дробно-дифференциальным оператором // Norwegian Journal of Development of the International Science. 2022. №99.С.10-15.