CBU Coding Challenge

MixData

Duch kelgan muammolar

Ma'lumotlarning turliligi

Ma'lumotlar turli fayllarda va bir-biriga qanday bog'lik ekanligini topish muammosi.

Data missing

Ma'lumotlar yetishmasligi yoki ortiqcha ekanligi, bir xil formatda emasligi.

Feature selection

60 dan ortiq ustunlar va ularning natijaga qanday darajada ta'sir qilishini aniqlash.

MixData

Muammolarni hal qilish yo'llari

Ma'lumotlarni birlashtirish

Turli faylarni o'zro birlashtiruvchi ustunni aniqlash va shu yordamida ma'lumotlarni bitta yaxlit ma'lumotlar to'plamiga keltirish.

Ma'lumotlar to'liqligi va to'g'riligini taminlash

Ma'lumotlarni ta'sir qilish darajasi va ma'nosiga qarab yetishmagan ma'lumotlarni to'ldirish, bir xil ma'lumotlar turiga olib kelish, ta'sir darajasi quyi bo'lgan ma'lumotlarni tozalash.

Korralatsiyon ta'sirni aniqlash

Hamma ustunlar bo'yicha korralatsiya darajasini aniqlash va kerakli ustunlarni tanlab olish, bazi ustunlar yordamida korralatsiya darajasi yuqori yangi ustunlarni yaratish

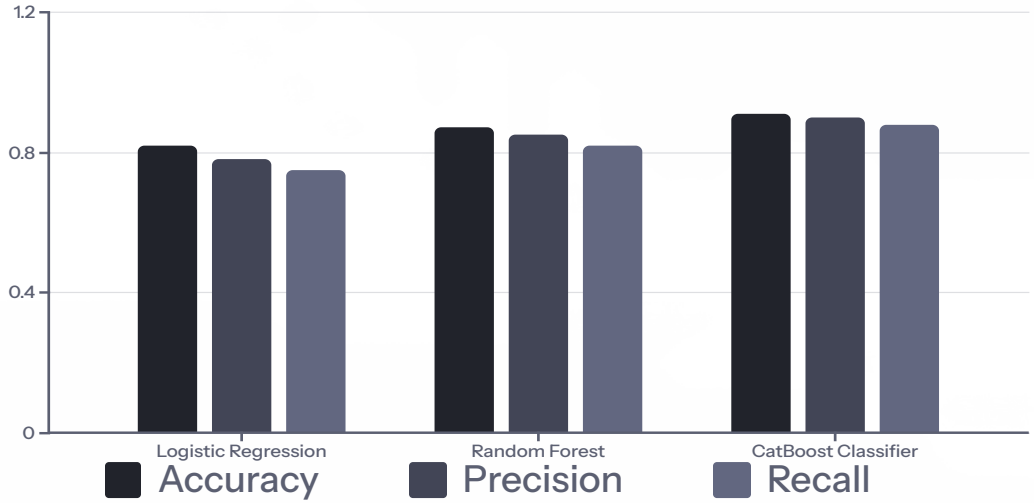
MixData

Modellarni o'qitish va muqobil modelni tanlash

Biz eng samarali kredit ballari modelini aniqlash maqsadida uchta asosiy algoritmni o'qitdik va taqqosladik:

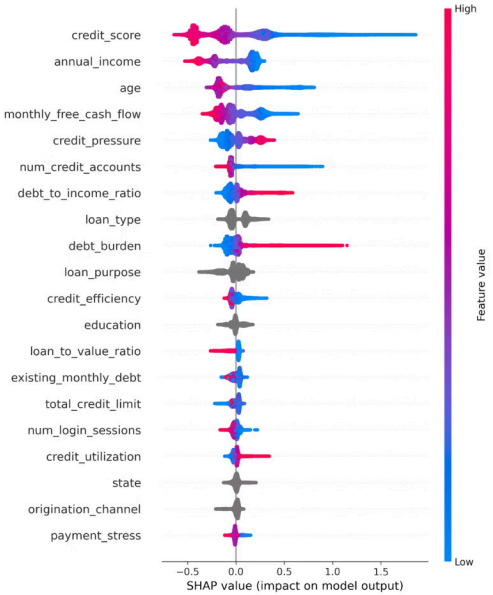
- **Logistic Regression**
- **Random Forest**
- **CatBoost Classifier**

Har bir model ma'lumotlar to'plamimizda sinovdan o'tkazildi va ularning ishlash ko'rsatkichlari, aniqligi va mustahkamligi tahlil qilindi. Bu taqqoslash natijasida eng yuqori baholash sifatini ta'minlaydigan optimal model tanlandi.



MixData

Tanlangan model arxitekturasi: CatBoost Classifier



Ma'lumotlarimizni tahlil qilish va modellashtirish uchun biz CatBoost klassifikatorini tanladik. Bu qaror ma'lumotlar to'plamimizning o'ziga xos xususiyatlari va CatBoost'ning afzalliklariga asoslangan:

1. **Kategorik xususiyatlarni avtomatik qayta ishlash**: Bizning ma'lumotlar to'plamimizda (masalan, `employment_type`, `education`, `loan_type`, `referral_code` kabi) juda ko'p kategorik xususiyatlar mavjud. CatBoost kategorik ma'lumotlarni oldindan kodlash (one-hot encoding) kabi murakkab ishlov berishlarsiz samarali boshqarish uchun maxsus yaratilgan. Bu ayniqsa `referral_code` kabi yuqori kardinalli xususiyatlar uchun juda foydali.
2. **Anomaliyalarga chidamlilik**: Biz raqamli xususiyatlardagi (`num_login_sessions`, `num_customer_service_calls` kabi) ba'zi anomaliyalarni aniqladik. CatBoost kabi daraxtga asoslangan modellar chiziqli modellarga nisbatan anomaliyalarga kamroq sezgir.
3. **Nomutanosib maqsadli o'zgaruvchi bilan ishlash**: `default` (defolt) ustunida sinflar nomutanosibligini (taxminan 5.1% defolt holatlari) qayd etdik. CatBoost nomutanosib ma'lumotlar to'plamlarini boshqarish uchun `class_weights` kabi parametrlarga ega, bu esa ozchilik sinfidagi ishlashni yaxshilashi mumkin.
4. **Yuqori samaradorlik**: CatBoost yuqori aniqlik va tezlik bilan mashhur bo'lib, ko'pincha boshqa gradient kuchaytiruvchi algoritmlardan (masalan, XGBoost va LightGBM) ustun turadi, ayniqsa raqamli va kategorik xususiyatlar aralashmasi bo'lgan ma'lumotlar to'plamlari bilan.
5. **Kamroq giperparametr sozlash ehtiyoji**: U ko'pincha sukutdagi parametrlar bilan yaxshi ishlaydi, bu esa giperparametr optimallashtirishga bo'lgan ehtiyojni kamaytiradi, garchi aniq sozlash natijalarni yanada yaxshilashi mumkin.

Yaratilgan model

Afzalliklari

- **Yuqori aniqlik:** Murakkab munosabatlarni aniqlay oladi.
- **Optimallashtirilgan:** Katta ma'lumotlar bilan samarali ishlaydi.
- **Moslashuvchan:** Turli vazifalarga moslashish imkoniyati.
- **O'zaro bog'liqlikni aniqlash:** Xususiyatlar orasidagi nozik aloqalarni topa oladi.

Kamchiliklari

- **Murakkablik:** Tushunish va sozlash biroz qiyinroq.
- **Manba talabchan:** Yuqori hisoblash resurslarini talab qilishi mumkin.
- **Izohlilik:** Ba'zan natijalarni sharhlash qiyin bo'lishi mumkin.

MixData