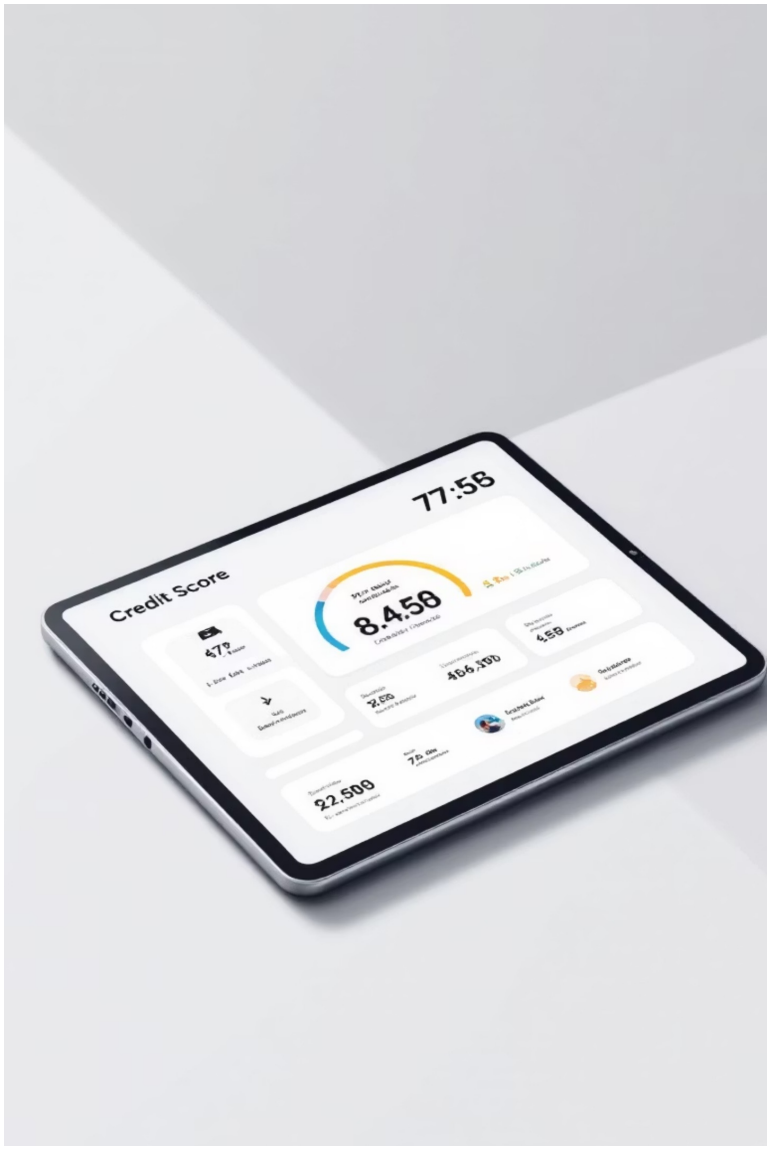



# CBU Coding Challenge

*Project: Scoring*

Group: Mix Data

# Duch kelgan muammolar

## Ma'lumotlarning turliligi

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

## EDA muammolari

1. Defolt xolatlari juda kamligi
2. Ba’zi ma’lumotlar yo’qligi
3. Ustunlar xar hil nom bilan takrorlangaligi va multikollinearlik

## Feature selection

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

Group: Mix Data

# 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

Group: Mix Data

# 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**

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.

Group: Mix Data

Tanlangan model arxitekturasi:
CatBoost Classifier

- Kategorik xususiyatlarni avtomatik qayta ishlash
- Anomaliyalarga chidamlik.
- Nomutanosib maqsadli o’zgaruvchiga moslasha olish
- Yuqori samaradorlik
- AUC: ~0.81 (cv)

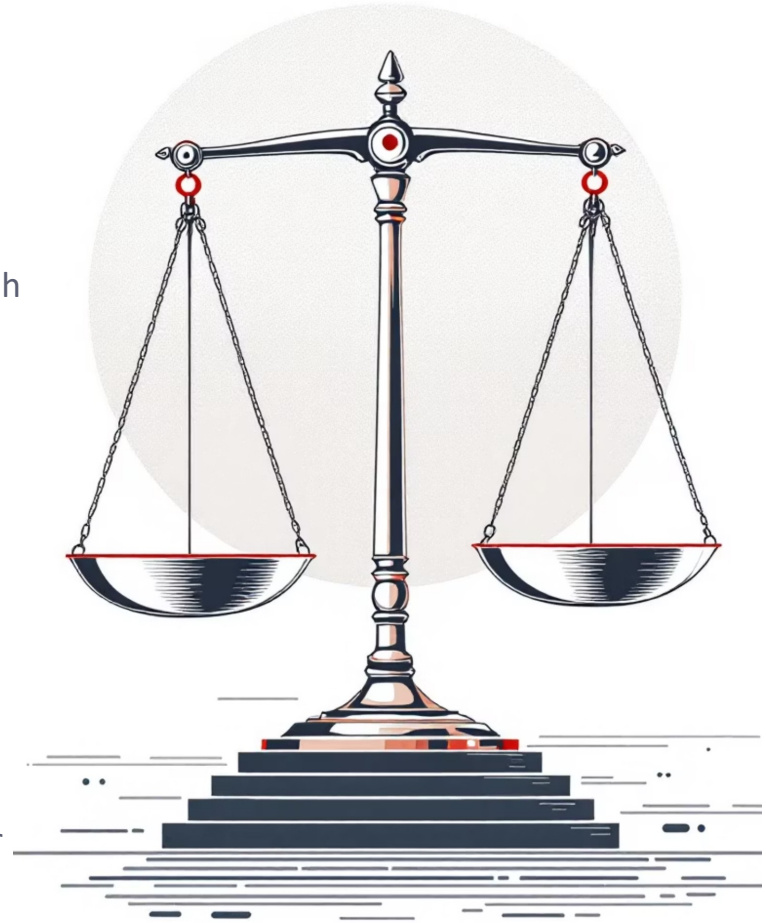
# Yaratilgan model

## Afzalliklari

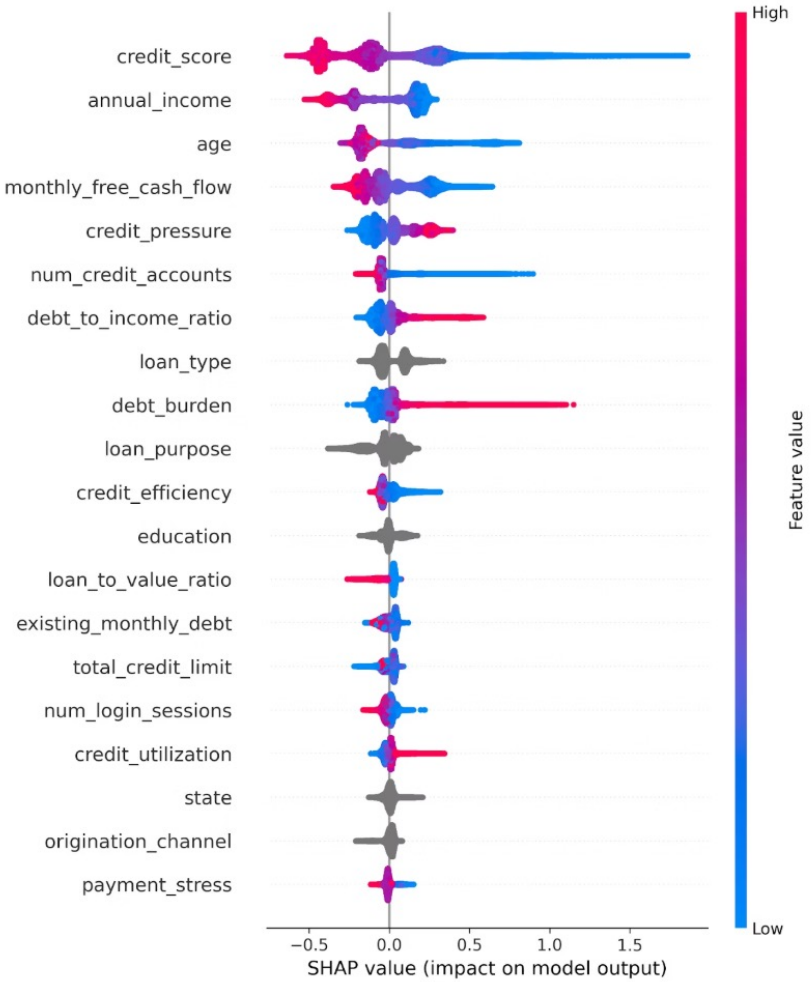
- **Yuqori aniqlik:** Murakkab munosabatlarni aniqlay oladi.
- **Optimallashtirilgan:** Katta ma'lumotlar bilan samarali ishlaydi.
- **Moslashuvchan:** Turli vazifalarga moslashish imkoniyati. Hamda turli xildagi datalar uchun qulay
- **O'zaro bog'liqlikni aniqlash:** Xususiyatlar orasidagi nozik aloqalarni 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.
- **Giperparametr topish:** Optuna hamda qo’lda giperparameterlar sozlangan.



Group: Mix Data




# SHAP

Top influencing parameters

- **Credit score.** Kichikroq bo’lsa defolt riski katta;
- **Income.** Kattaroq bo’lsa defolt riski kichrayadi;
- **Age.** Kattalar to’lay olmaydigan kreditni kamroq olishadi;
- **Credit pressure (hisoblangan qiymat).** Kamroq bo’lsa kredit to’lash osonroq;

E’tiboringiz uchun
Rahmat