HOME CREDIT DEFAULT RISK

1.1 RAZUMEVANJE PROBLEMA

Mnogi ljudi širom sveta imaju poteškoće da dobiju kredit zbog nedovoljno razvijene ili nepostojeće kreditne istorije. Neke organizacije često postaju meta nepouzdanih kreditora, što ih dovodi u dodatni rizik. Home Credit nastoji da proširi finansijsku inkluziju tako što pruža sigurno i transparentno iskustvo zaduživanja, koristeći različite izvore alternativnih podataka (telekomunikacione zapise, transakcione podatke i sl.) kako bi procenila sposobnost klijenata da otplaćuju kredite.

Problem koji se razmatra u ovom radu jeste predikcija da li će klijent uspeti da otplati kredit ili ne, što predstavlja zadatak binarne klasifikacije. Ovaj problem je od velikog značaja, jer banke i finansijske institucije moraju da minimizuju rizik od nenaplativih kredita, a istovremeno da omoguće pristup finansijskim uslugama što širem krugu klijenata.

Cilj projekta je razvoj prediktivnog modela koji će na osnovu raspoloživih podataka umeti da razlikuje rizične od sigurnih klijenata, čime se omogućava donošenje boljih poslovnih odluka i unapređuje proces kreditnog odlučivanja.

1.2 RAZUMEVANJE PODATAKA

1.2.1 OPIS DATASET-A

application_train / application_test
Glavna tabela sa statičkim podacima o zahtevima za kredit. Jedan red = jedan kredit (aplikacija). Train sadrži ciljnu promenljivu TARGET, test ne sadrži. Ključ za spajanje sa ostalim tabelama je SK_ID_CURR.

bureau
Svi prethodni krediti klijenta kod drugih finansijskih institucija koje je prijavio kreditni biro. Za svaku tekuću aplikaciju u uzorku ovde se pojavljuje onoliko redova koliko klijent ima prethodnih kredita u birou. Veza sa aplikacijom: SK_ID_CURR; veza sa mesečnim stanjem: SK_ID_BUREAU.

bureau_balance
Mesečna stanja prethodnih kredita iz kreditnog biroa. Jedan red = jedan mesec istorije za jedan prethodni kredit (tabela ima mnogo redova). Veza: SK_ID_BUREAU.

POS_CASH_balance
Mesečni snapshot-i prethodnih POS i “cash” kredita koje je klijent imao u Home Credit-u. Jedan red = jedan mesec istorije za jedan prethodni HC kredit (consumer/cash). Veza: SK_ID_PREV.

credit_card_balance
Mesečni snapshot-i prethodnih kreditnih kartica u Home Credit-u. Jedan red = jedan mesec istorije za jednu karticu. Veza: SK_ID_PREV.

previous_application
Sve prethodne prijave za kredit u Home Credit-u za klijente koji su u našem uzorku. Jedan red = jedna ranija prijava. Veza sa glavnom tabelom preko SK_ID_CURR; sekundarni ključ za povezivanje sa mesečnim tabelama je SK_ID_PREV.

installments_payments
Istorija otplata za ranije odobrene HC kredite (vezane za naše klijente). Jedan red = jedna uplaćena rata ili propuštena rata. Veza: SK_ID_PREV.

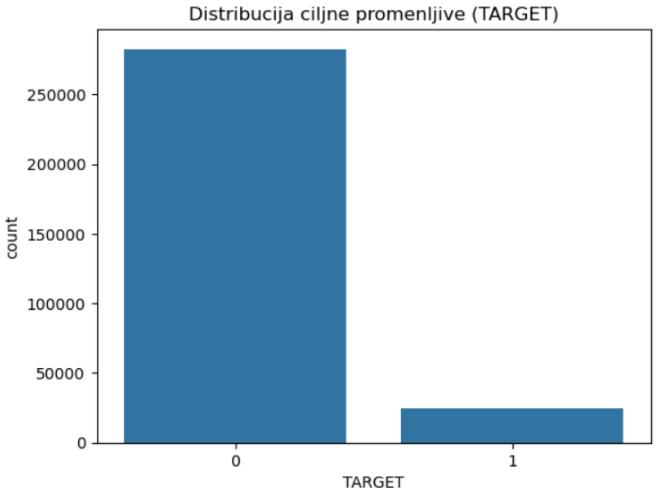
Zbog odnosa "jedan-prema-više" (npr. jedan SK_ID_CURR <-> više SK_ID_BUREAU/ SK_ID_PREV) pomoćne tabele se tipično agregiraju po SK_ID_CURR (mean, sum, max, count...) pre spajanja sa glavnom tabelom.

1.2.2 CILJNA PROMENLJIVA I DISTRIBUCIJA KLASA

Cilj ovog projekta je da se razvije model mašinskog učenja koji predviđa verovatnoću da klijent ne otplati kredit (promenljiva TARGET: 1 = klijent je u problemu sa otplatom, 0 = klijent uredno vraća kredit). Ovaj problem je klasičan primer binarne klasifikacije sa neuravnoteženim klasama - daleko više je klijenata koji vraćaju kredit nego onih koji ne vraćaju.

U train skupu posle početnog čišćenja ima približno 307 507 primera, pri čemu je udeo klijenta sa TARGET= 1 oko 8.07%, a sa TARGET=0 oko 91.93%. To jasno ukazuje na izraženu neravnotežu klasa.

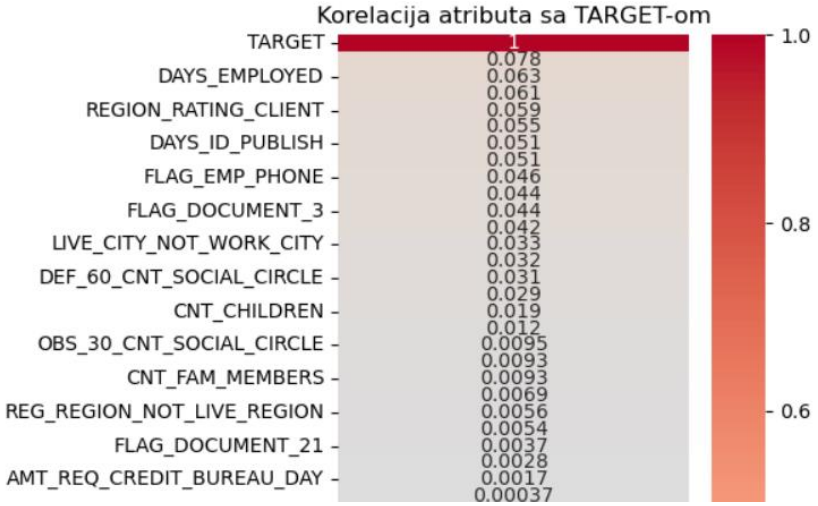
Zbog toga sama tačnost (accuracy) nije dovoljna mera uspeha, već se naglasak stavlja na ROC-AUC metriku, jer pokazuje kako se ponaša model kada se menja prag (threshold) tj. koliko je dobar u razlikovanju klasa.



Da bi se umanjio efekat neravnoteže, primenjena je stratifikovana podela i cross-validacija (StratifiedKFold), čime je obezbeđeno da odnos klasa ostane stabilan kroz sve foldove i da procena generalizacije bude pouzdanija. Uz to, kod logističke regresije i Random Forest-a korišćen je parametar class_weight="balanced", kako bi se greške na manjinskoj klasi teže kažnjavale i kako bi model bio osetljiviji na primere neotplate. Konačno, prag odluke je dodatno podešen optimizacijom prema F1 meri, pri čemu se pokazalo da je najbolji prag oko 0,161.

1.3 PRIPREMA PODATAKA

Prvo je rešen problem nedostajućih vrednosti i anomalija. Vrednost 365243 u koloni DAYS_EMPLOYED tretirana je kao nerealan sentinel i zamenjena je sa NaN. Potom je izračunat procenat NaN po koloni i uklonjene su kolone sa više od 50% nedostajućih vrednosti, čime je smanjen šum i rizik od pristrasne imputacije. Za preostale numeričke atribute primenjena je imputacija medijanom (SimpleImputer), a za kategorijske atribute imputacija najčešćom vrednošću. Zatim je primenjen low variance filter tj. izbacivanje kolona sa malom varijansom vrednosti, jer one ne pomažu modelu već mogu dodati šum, praktično su konstantne. Nakon toga su izbačene visoko-korelisane kolone gde se kao prag za visoku korelaciju koristio threshold=0.7. Nakon kasnijih spajanja tabela, preostale praznine u agregatnim kolonama popunjene su neutralnom nulom (fillna(0)) kako bi se obezbedila konzistentnost dizajn-matrice.



Kod kategorijskih promenljivih primenjen je, na udruženom skupu (train ∪ test), one-hot encoding (get_dummies, drop_first=True). Ovim pristupom obezbeđeno je da train i test dele identičan skup kolona i da ne dođe do nesklada atributa prilikom modelovanja.

Pre spajanja sa pomoćnim tabelama rađena je agregacija zbog odnosa jedan-prema-više u pomoćnim tabelama. Za kreditni biro (bureau) najpre su agregirana mesečna stanja (bureau_balance -> bb_agg po SK_ID_BUREAU), pa je rezultat vraćen u bureau i dalje agregiran na nivo klijenta (po SK_ID_CURR; npr. DAYS_CREDIT [mean/max/min], AMT_CREDIT_SUM [sum/mean/max]...). U okviru Home Credit istorije prethodnih zahteva, previous_application je obogaćen u više koraka:

(1) installments_payments je pretvoren u pokazatelje kašnjenja i discipline plaćanja

LATE_DAYS = DAYS_ENTRY_PAYMENT – DAYS_INSTALMENT,

PAYMENT_RATIO = AMT_PAYMENT / AMT_INSTALMENT,

pa agregiran po SK_ID_PREV;

(2) POS_CASH_balance je agregirao „days past due“ metrike (SK_DPD, SK_DPD_DEF) i dužinu istorije;

(3) credit_card_balance je dao pokazatelje zaduženosti kroz

UTILIZATION = AMT_BALANCE / AMT_CREDIT_LIMIT_ACTUAL.

Ovi skupovi su spojeni nazad na previous_application, a zatim su numerički agregati podignuti na nivo SK_ID_CURR. Paralelno su konstruisani i kategorijski agregati (get_dummies + groupby sum/mean) za najreprezentativnije kategorije u bureau i previous_application, kako bi se informacija o učestalosti i udelima kategorija sačuvala u vektorskom obliku pogodnom za modele. Konačno, kompletni agregati su pridruženi glavnoj tabeli aplikacija (X_final i test_final).

U okviru „feature engineering"-a dodati su deskriptivni i odnosni indikatori koji hvataju starost i stabilnost klijenta, kao i opterećenje budžeta: AGE (−DAYS_BIRTH/365), EMPLOYED_YEARS (−DAYS_EMPLOYED/365), EMPLOYED_AGE_RATIO (odnos radnog staža i starosti) i INCOME_PER_CHILD (prihod prilagođen broju dece). Iz transakcionih istorija dobijeni su i sažeti indikatori kašnjenja i zaduženosti (npr. INST_LATE_DAYS_, INST_PAYMENT_RATIO_, POS_SK_DPD_, CC_UTILIZATION_), koji su se pokazali korisnim u modelima zasnovanim na stablima.

Priprema za modele uključila je i skaliranje tamo gde je to metodološki važno. Za linearne modele (npr. logistička regresija) korišćen je StandardScaler (fit na train, transform na validation), dok su oni bazirani na stablima (Random Forest, XGBoost/LightGBM) trenirani nad neskaliranim, ali obogaćenim atributima, u skladu sa dobrom praksom za takve modele.

1.4 MODELOVANJE (IZBOR ALGORITAMA, TUNING)

U okviru modelovanja koristila sam ove algoritme: logistička regresija (baseline i referentni linearni model), Random Forest (ansambl stabala bez potrebe za skaliranjem), i XGBoost (gradijentni ansambl koji važi za veoma jak pristup na tabelarnim podacima sa nelinearnostima i interakcijama).

Za podešavanje hiperparametara korišćen je RandomizedSearchCV isključivo nad XGBoost-om, sa metrikom roc_auc i stratifikovanom podelom na 2 folda. U pretrazi su varirani n_estimators ∈ {200, 400, 600}, learning_rate [0.03, 0.05, 0.1], max_depth [3, 4, 5], subsample [0.8, 1.0], colsample_bytree [0.8, 1.0] i gamma [0, 0.1], dok je uključen rani prekid (early_stopping_rounds=50) uz fiksni eval_set=(X_val, y_val). Najbolja kombinacija je izabrana na osnovu prosečnog AUC-ROC rezultata kroz foldove tokom RandomizedSearchCV, a zatim su dobijeni best_params_ usvojeni za finalni XGBoost model, uz zadržavanje ranog prekida na istom validacionom skupu. Logistička regresija je korišćena kao baseline sa fiksnim podešavanjima (max_iter=2000, class_weight="balanced"), bez dodatnog tuninga C/penalty, dok je Random Forest treniran sa n_estimators=200, max_depth=10 i class_weight="balanced", bez dodatne pretrage hiperparametara.

Kao glavna metrika korišćen je AUC-ROC (zbog neravnoteže), uz izveštavanje F1 mere pri optimizovanom pragu odluke i precision-recall radi procene ponašanja nad manjinskom klasom. Nakon validacije, prag odluke je podešen na osnovu F1, čime je postignut bolji kompromis između precision i recall za slučajeve neotplate kredita.

1.5 EVALUACIJA (POREĐENJE PERFORMANSI)

1.5.1 REZULTATI PO MODELIMA

Za ocenu kvaliteta korišćene su metrike prilagođene neravnoteži klasa. Primarna metrika je AUC-ROC, jer meri sposobnost modela da razdvaja klase nezavisno od praga. Uz to su izračunati F1 i pregled Precision/Recall vrednosti kako bi se procenilo ponašanje na manjinskoj klasi (TARGET=1).

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.695181	0.165323	0.685599	0.266406	0.758270
Random Forest	0.752512	0.180817	0.585096	0.276259	0.746725
XGBoost (tuned)	0.920100	0.594796	0.032226	0.061139	0.774429

Pošto je XGBoost pri podrazumevanom pragu imao jako nizak recall (0.032), prag odluke je naknadno podešen prema F1. Najbolji prag je ≈ **0.1608**, pri čemu metrika postaje:

- Precision = **0.2682**
- Recall = **0.4175**
- F1 = **0.3266**

Ovo potvrđuje da je, kod nebalansiranih skupova, podešavanje praga ključno: AUC-ROC XGBoost-a je najviši (≈0.774), ali bez podešavanja praga model propušta većinu onih koji neće isplatiti kredit; pomeranjem praga postiže se bolji kompromis precision/recall.

1.5.2 FEATURE IMPORTANCE

Top 10 atributa koji najviše doprinose predviđanju modela:

- EXT_SOURCE_2 (0.1236)
- EXT_SOURCE_3 (0.1100)
- BUREAU_DAYS_CREDIT_MEAN (0.0365)
- DAYS_EMPLOYED (0.0297)
- EMPLOYED_YEARS (0.0292)
- BUREAU_CREDIT_ACTIVE_Closed_MEAN (0.0261)
- AGE (0.0235)
- DAYS_BIRTH (0.0232)
- PREV_INST_LATE_DAYS_MEAN_MEAN (0.0224)
- PREV_NAME_CONTRACT_STATUS_Refused_MEAN (0.0210)

1.5.3 EVALUACIJA

Po AUC-ROC, **XGBoost** je najuspešniji (≈0.774), što je očekivano na tabelarnim podacima sa nelinearnostima. Međutim, bez podešavanja praga daje nizak recall. To je logično jer je FN veliki a to je skuplja greška u poređenju sa FP. Posle optimizacije praga (≈0.161) F1 se značajno poboljšava uz razuman odnos precision/recall.

Logistička regresija daje solidan AUC (≈0.758) i vrlo visok recall na default pragu, ali uz nizak precision (visok FP).

Random Forest je između ova dva po većini metrika (AUC ≈0.747), uz uravnoteženije ponašanje od LR, ali slabiji AUC od XGBoost-a.

Najinformativnije promenljive dolaze iz eksternih skorova (EXT_SOURCE_*), „demografije“ (godine/starost, staž), i istorije iz biroa/prethodnih aplikacija - što je u skladu sa domenom kreditnog rizika.

1.6 PRIMENA MODELA

Model se može ugraditi u proces odobravanja kredita kao ,,risk scoring” komponenta koja svaku prijavu pretvara u verovatnoću da će klijent imati poteškoća sa isplatom (TARGET=1). Na osnovu te verovatnoće i izabranog praga (u mom radu prag ≈ 0,161 optimizovan za F1) banka može definisati tri akcije:

1. ,,Odobri” zona (score ispod donje granice): brza obrada bez dodatnih provera.
2. ,,Ručno proveri” zona (između donje i gornje granice): dodatna dokumentacija, telefonska verifikacija, provera prihoda ili kontakt poslodavca.
3. ,,Odbij” zona (iznad gornje granice): visok rizik da neće otplatiti kredit.

Pored binarne odluke, skor se može koristiti i za personalizaciju uslova: iznos kredita, rok otplate, kamatna stopa, minimalno učešće, ili obavezno dodatno osiguranje - što smanjuje kreditni rizik uz očuvanje dostupnosti kredita.

1.7 ZAKLJUČAK

Najbolji rezultat u ovom radu postiže **XGBoost** sa **AUC-ROC ≈ 0,774**. Nakon podešavanja praga odluke na ≈ 0,161 postiže se uravnoteženiji odnos za manjinsku klasu (povećan recall uz prihvatljiv precision), što je poslovno poželjno kada je cilj pravovremeno identifikovati rizične klijente. Logistička regresija je poslužila kao stabilna polazna tačka (baseline), dok je Random Forest dao solidne rezultate bez zahtevnog tuninga.

Moguća unapređenja:

(1) dodatni feature engineering na istorijskim tabelama da bi se izvuklo još indikatora koji prikazuju ponašanje kroz vreme

(2) finiji cost-sensitive prag (optimizacija prema poslovnim troškovima FP/FN, ne samo F1), usmesto da se prag bira po F1, izračuna se očekivani trošak = FP_cost × FP + FN_cost × FN i

nađe se prag koji taj trošak minimizuje. Tako se dobija poslovno optimalna tačka koja odgovara svakoj banci.

(3) dublja interpretacija (npr. SHAP pokazuje koliko je svaka osobina „gurala“ skor ka višem ili nižem riziku na nivou pojedinačnog klijenta i globalno).

(4) stroža validacija i monitoring u produkciji (npr. stabilniji cv).

Očekivani efekat ovih koraka je dodatno podizanje AUC-a i, važnije, bolji kompromis između profita i rizika u realnom kreditnom toku.