Mozgalo: Predviđanje ponašanja klijenata

Ana Parać
Prirodoslovno matematički fakultet
Sveučilište u Zagrebu
Zagreb, Hrvatska
anaparac8@gmail.com

Sara Pužar
Prirodoslovno matematički fakultet
Sveučilište u Zagrebu
Pula, Hrvatska
sara.puzar@gmail.com

Paula Vujasić
Prirodoslovno matematički fakultet
Sveučilište u Zagrebu
Zagreb, Hrvatska
paula.vujasic@gmail.com

***Sažetak*—Predstavljamo rješenje nagradnog zadatka danog na natjecanju Mozgalo, pod nazivom "Client Behavior Prediction". Cilj zadatka bio je predvidjeti hoće li određeni ugovor s bankom biti prijevremeno prekinut ili ne. Naše konačno rješenje ostvareno je pomoću XGBoost metode.**

***Index Terms*—strojno učenje, previđanje ponašanja, Mozgalo**

I. Uvod

Predviđanje ponašanja klijenata vrlo je važno u bankarskom sektoru kako bi se smanjio sveukupni rizik te varijabilnost poslovanja. Banke nastoje upravljati rizicima predviđajući ponašanje klijenata u ovisnosti o različitim faktorima, koji mogu biti osobnog karaktera ili makroekonomski pokazatelji. Poznavanjem utjecanja vanjskih faktora na ponašanje klijenata banke nastoje poboljšati svoje usluge pripremom novih ponuda.

Cilj ovog projekta je izrada binarnog klasifikatora koji predviđa hoće li ugovor s bankom biti prijevremeno raskinut. Zadatak je dan na natjecanju Mozgalo te su ga osmislili zlatni partneri natjecanja, Adacta i Raiffeisen Bank.

II. Opis problema

A. Zadatak

Zadatak je predvidjeti klijente koji će promijeniti ugovoreni odnos s bankom (kredit ili depozit) tako da prijevremeno raskinu ugovor. S obzirom na dobiveni dataset s labeliranim podacima, rješavanju problema pristupile smo kao problemu nadziranog učenja. Zadatak je postavljen kao problem binarne klasifikacije, gdje jedna klasa predstavlja prijevremeno prekinute ugovore, a druga ugovore koji nisu prijevremeno prekinuti.

B. Skup podataka

Skup podataka za rješavanje zadatka osigurala je RBA. Podatci sadrže informacije od 2010. godine, o kreditima i depozitima ugovorenim s bankom. Dane su sljedeće informacije:

- datum izvještavanja (banka generira izvještaje za svaki kvartal)
- ID klijenta
- oznaka partije (jedinstvena oznaka nekog ugovora izmedu banke i klijenta)
- datum otvaranja ugovora
- planirani datum zatvaranja
- stvarni datum zatvaranja ugovora
- ugovoreni iznos
- stanje na kraju prethodnog kvartala
- stanje na kraju trenutnog kvartala
- valuta
- vrsta klijenta
- proizvod (šifra proizvoda)
- vrsta proizvoda (kredit ili depozit)
- visina kamate
- tip kamate
- starost (starost pravne ili fizičke osobe)
- prijevremeni raskid (da ili ne)

Skup smo procesirale na sljedeći način:

PRVA FAZA:

- promjena vrijednosti prijevremenog raskida prema uputama organizatora (ugovor mora biti isplaćen najmanje 10 dana ranije da bi se smatrao prijevremeno raskinutim)
- uklonjeni retci u kojima ja datum otvaranja veći od planiranog datuma zatvaranja
- uklonjeni retci u kojima je ugovoreni iznos 0
- uklonjeni retci u kojima nije specificiran planirani datum zatvaranja

DRUGA FAZA:

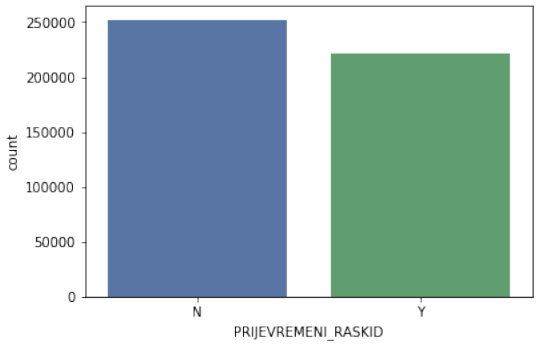
- za svaki ugovor ostavljen je jedan izvještaj – onaj u kojem se prvi put spominje datum zatvaranja
- za ugovore za koje ne postoji datum zatvaranja, uzeti su zadnji izvještaji ukoliko je zadnji datum izvještavanja manji od planiranog datuma zatvaranja (kako samo za te ugovore možemo sa sigurnošću reći da nisu prijevremeno otplaćeni)

TREĆA FAZA:

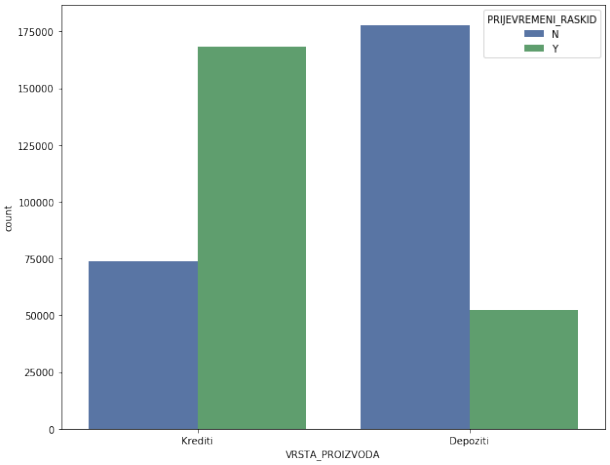
- dodavanje osmišljenih značajki
- u retcima gdje je starost negativna ili veća od 900, starost je postavljena na nedefiniranu vrijednost (Nan)

- u retcima gdje je visina kamate nedefinirana i gdje je jednaka 72, visina kamate je postavljena na nedefiniranu vrijednost (Nan)

Nakon procesiranja skupa podataka, ostvarena je distribucija Y: 46.69 - N: 53.3, dakle dobiveni skup je balansiran.



Slika 1. Distribucija prijevremeno i neprijevremeno prekinutih ugovora



Slika 2. Distribucija prijevremenog raskida ugovora po kreditima i depozitima

Podatke smo vizualizirale pomoću histograma u kojima prikazujemo utjecaj ugovorenog iznosa, starosti i visina kamate. Histogram u kojem prikazujemo visinu kamate očekivan je s obzirom na distribuciju kredita i distibuciju depozita na skupu podataka kojeg smo dobile nakon pretprocesiranja. Naime, krediti su većinski prijevremeno otplaćeni, dok depoziti uglavnom nisu prijevremeno prekinuti. Kako depoziti imaju bitno niže kamate od kredita, ugovori s višim kamatama su češće prijevremeno otplaćeni. Prikaz odnosa starosti i prijevremenog otplaćivanja daje naslutiti kako stariji klijenti imaju tendenciju ne prekidati prijevremeno ugovore.

III. Metoda

A. Značajke

Rješavanje problema započele smo kreiranjem novih značajki. Navodimo one koje smo koristile u konačnom modelu.

- Povijest prijevremenih - broj prijevremeno prekinutih ugovora nekog klijenta
- Povijest neprijevremenih - broj ugovora nekog klijenta koji nisu prijevremeno prekinuti
- Nadvrsta klijenta - podjela klijenata na fizičke i pravne osobe
- Mjeseci - procijenjeno vrijeme za planirano trajanje ugovora izraženo u mjesecima
- Jednodnevni - ugovori kojima je datum otvaranja jednak planiranom datumu zatvaranja
- Zatvaranje - godina planiranog zatvaranja ugovora
- Otvaranje - godina otvaranja ugovora
- GDP - bruto domaći proizvod u godini otvaranja ugovora
- Prosječna plaća - prosječna plaća u godini otvaranja ugovora
- Nezaposlenost - stopa nezaposlenosti u godini otvaranja ugovora
- Price indeks - PPI u godini otvaranja ugovora

Uz navedene značajke u konačnom modelu koristile smo i neke od informacija dobivenih u skupu podataka: vrsta proizvoda (kredit - 0, depozit - 1), visina kamate i neke od značajki nastalih one hot encoding-om valute, tipa kamate, vrste proizvoda i vrste klijenta. Sve makroekonomske čimbenike vezale smo uz godinu otvaranja jer je svrha zadatka predvidjeti prijevremeni prekid u samom trenutku otvaranja ugovora. Osim navedenih značajki kreirale smo i sljedeće: tečaj eura, franka i dolara u trenutku otvaranja ugovora, bruto domaći proizvod po glavi stanovnika, kamate nula (kategorička varijabla koja govori je li visina kamate jednaka nula), no nismo ih koristile u konačnom rješenju. 25 značajki koje smo odabrale za konačno rješenje izdvojile smo na osnovi važnosti prema kriteriju dobiti (eng. gain), koja se izračunava na temelju doprinosa svake značajke za svako stablo u modelu. Izdvojene značajke pokrivaju 95% varijabilnosti. Pokušale smo odrediti značajke i prema drugim kriterijima, no za dobit je ostvarena najbolja točnost. U tablici (5) u Dodatku prikazujemo koreliranost među odabranim značajkama.

B. Model

Problem smo pokušale riješiti pomoću više metoda. Prvi pokušaj bio je pomoću logističke regresije. Budući da je u skupu podataka bilo mnogo nedostajućih vrijednosti, te da je kod logističke regresije potrebno ručno stvaranje interakcija između značajki, odlučile smo pokušati s Random Forest-om te kombiniranjem te dvije metode ovisno o podjeli na depozite i kredite. Navedene metode isprobavale smo za vrijeme trajanja natjecanja.

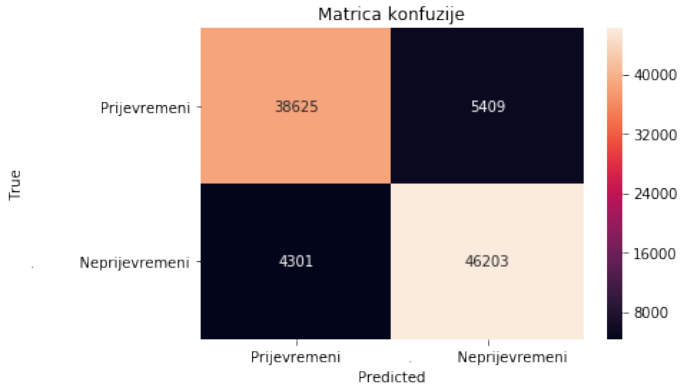
Nakon završetka natjecanja nastavile smo rad na zadatku te smo se odlučile za XGBoost metodu kao konačno rješenje. Izbor hiperparametara provele smo unakrsnom validacijom (5-fold CV).

IV. Prikaz rezultata

Dobiveni skup podataka podijelile smo na sljedeći način. 80% dobivenog skupa odvojile smo za treniranje i 20% za testiranje. Na dijelu za treniranje radile smo unakrsnu validaciju (5-fold CV) kako bismo odredile najbolje parametre za XGBoost.

Skup	**Točnost (*Accuracy*)**	**F1 mjera**
Unakrsna validacija	89.61% (*prosjek*)	88.75% (*prosjek*)
Testni skup	89.72%	88.83%

Slika 3. Rezultati treniranja i testiranja



Slika 4. Matrica konfuzije na rezultatima test seta

Na testnom skupu Mozgala, na kojem smo za vrijeme trajanja natjecanja nekoliko puta dnevno imale priliku provjeravati točnost modela, postigle smo točnost od 70% i F1 mjeru od 73%. Na validacijskom skupu Mozgala, za kojeg smo imale pravo samo jednom objaviti rješenje, postigle smo točnost od 69% i F1 mjeru od 69%.

V. Osvrt na druge pristupe

Problem prijevremene otplate kredita puno je manje istraživan od problema klasifikacije na tzv. dobre i loše kredite (loši su krediti oni koji neće biti otplaćeni). Malekipirbazarii Aksakalli (2015.) su problem neotplaćenih kredita riješili pomoću Random Forest metode te rezultate usporedili s rezultatima dobivenim pomoću metode potpornih vektora, logističke regresije te pomoću K-najbližih susjeda.

Novija istraživanja koja se bave prijevremenom otplatom kredita uglavnom se odnose na tzv. online kredite. Tako su Zhiyong Li, Ke Li, Xiao Yao i Qing Wen (2019.) problem riješili korištenjem multivarijantne logističke regresije. U svom su radu integrirali makroekonomske pokazatelje i podatke o kreditima. U starijem istraživanju Goodarzi, Kohavi, Harmon i Senkut (1998.) koristili su naivni Bayesov klasifikator.

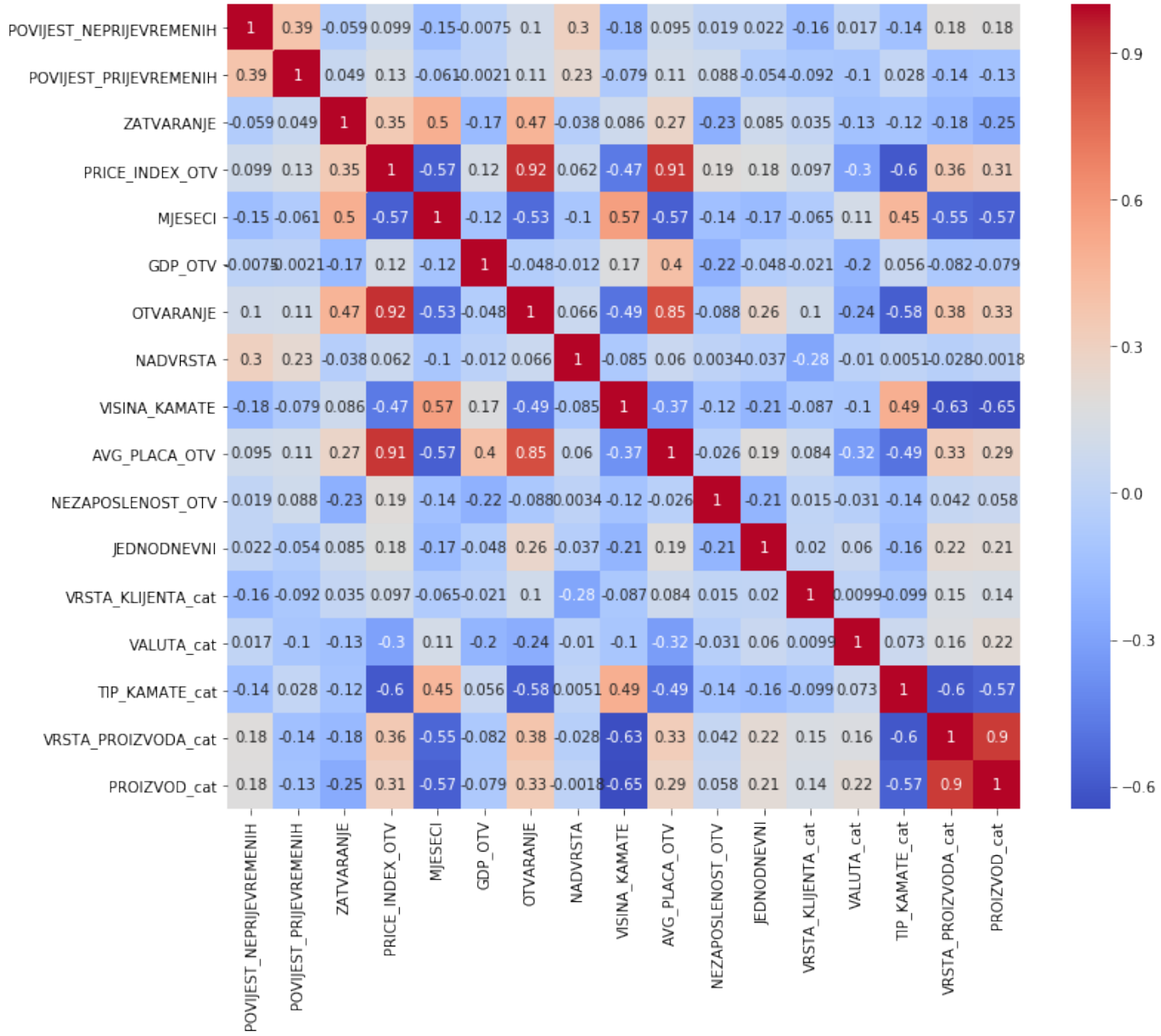
U nastavku navodimo neka od rješenja finalista natjecanja Mozgalo. Pobjednički je pristup bio ansambl nekliko različitih metoda, primjerice XGBoosta, Random Forest metode, KNN-a, neuronskih mreža i drugih dobiven pomoću stacking-a. Jedan od timova problemu je pristupio kao problemu regresije, procijenjujući koliko ranije će se ugovor s bankom prekinuti. Finalisti su također koristili i Random Forest metodu i XGBoost.

VI. Mogući budući nastavak istraživanja

U budućem razvoju rješenja isprobale bismo ansambl nekoliko različitih metoda, primjerice XGBoosta, Random forest metode i logističke regresije. Za daljnju razradu predlažemo detaljniji razvoj modela podijeljenog na fizičke i pravne osobe, s različitim značajkama. S obzirom na distribuciju fizičkih i pravnih osoba u datasetu model za fizičke osobe učio bi samo na podacima o fizičkim osobama, a pravne na svima. Pokušale bismo kreirati još nekoliko značajki koje se odnose na ”povijest” klijenata, primjerice uzimajući u obzir omjer prijevremeno prekinutih ugovora naspram ukupnog broja ugovora određenog klijenta. Također bismo, uz savjetovanje eksperata na području financija, pokušale uklopiti i neke druge makroekonomske čimbenike.

Literatura

[1] Machine Learning @Coursera: Stanford; Washington
[2] radionice natjecanja Mozgalo
[3] Zhiyong Li, Ke Li, Xiao Yao, Qing Wen: Predicting Prepayment and Default Risks of Unsecured Consumer Loans in Online Lending, 2019.
[4] Afshin Goodarzi, Ron Kohavi, Richard Harmon, Aydin Senkut: Loan Prepayment Modeling, 1998.
[5] Milad Malekipirbazari, Vural Aksakalli: Risk assessment in social lending via random forests, 2015.



Slika 5. Tablica koreliranosti izabranih značajki