# Mozgalo - Client Behavior Prediction: A Machine Learning Challenge

Elena Murljačić
*Prirodoslovno-matematički fakultet*
*Sveučilište u Zagrebu*
e.murljacic11@gmail.com

Dorotea Rajšel
*Prirodoslovno-matematički fakultet*
*Sveučilište u Zagrebu*
dorotea.rajsel@gmail.com

Darija Strmečki
*Prirodoslovno-matematički fakultet*
*Sveučilište u Zagrebu*
darija.strmecki@gmail.com

Petra Vlaić
*Prirodoslovno-matematički fakultet*
*Sveučilište u Zagrebu*
petra.vlaic@gmail.com

***Sažetak*—U ovom radu, rješavali smo klasifikacijski zadatak dan u sklopu studentskog natjecanja Mozgalo. Cilj problema je pokušati predvidjeti klijente koji će potencijalno promijeniti ugovoreni odnos s bankom u smislu prijevremene otplate kredita ili produženja oročenog depozita. Modeli koji su postizali najbolje rezultate bili su dobiveni koristeći XGBoost algoritam te slučajne šume. Za implementaciju rješenja, dodatno smo isprobali i neuronske mreže. Konačno, izabran je XGBoost algoritam uz pretraživanje kojim je dobivena točnost od 0.78 i F1 score 0.75 (na danom skupu za testiranje).**

***Index Terms*—strojno učenje, PE format, XGBoost, slučajne šume, neuronske mreže**

## I. Uvod

Pri sklapanju ugovora između banke i klijenta, banka razmatra razne faktore koji bi mogli utjecati na ponašanje klijenta. Ono ovisi o internim i eksternim faktorima kao što su npr. demografija (starost klijenata, broj djece, bračni status), makroekonomski podatci (snaga ekonomije, stopa zaposlenosti, BDP) te naravno visina kamatnih stopa na tržištu. Problem je predvidjeti rizično ponašanje klijenata.

Prijevremena otplata kredita ili produženje oročenog depozita potencijalne su promjene ugovorenog odnosa koje banka pokušava predvidjeti. Uzimajući u obzir ponašanje klijenata kroz duži vremenski period, kreiraju se nove personalizirane ponude. U konačnici, cilj projekta je umanjiti negativne efekte koje takvo rizično ponašanje može imati na financijski rezultat banke.

## II. Opis problema

Na raspolaganju su nam podaci o transakcijskim računima pravnih osoba i/ili stambenih kredita fizičkih osoba u nekoliko vremenskih nizova. Podatke o transakcijskim računima pravnih i fizičkih osoba ustupila je RBA. Za bolju klasifikaciju, iskoristili smo podatke o makroekonomskim pokazateljima sa stranica Državnog zavoda za statistiku [1], RBA [2] i Hrvatske narodne banke [3] (kao što su prosječna plaća, BDP, inflacija, državni dug i stopa zaposlenosti).

### A. Analiza podataka

Skup podataka sastoji se od 5193124 redaka (zapisa) i 17 stupaca (značajki). Te značajke su: datum izvještavanja, ID klijenta, oznaka partije, datum otvaranja, planirani datum zatvaranja, datum zatvaranja, ugovoreni iznos, stanje na kraju prethodnog kvartala, stanje na kraju kvartala, valuta, vrsta klijenta, proizvod, vrsta proizvoda, visina kamate, tip kamate, starost klijenta i prijevremeni raskid, pri čemu je prijevremeni raskid ciljna značajka.

Problem je klasifikacijski: ciljna značajka je kategorička varijabla koja govori je li ugovor (kredit ili štednja) prijevremeno raskinut ili ne. U našem skupu za treniranje otprilike 70% ugovora je pravovremeno raskinuto, dok je 30% raskinuto prijevremeno.

### B. Procesiranje podataka

Kao prvi korak iz skupa podataka izbrisali smo izvještaje kod kojih jedna partija pripada više osoba te smo na taj način dobili 99.89% početnog skupa podataka. Također, prema preporuci Mozgalo tima, u stupcu 'prijevremeni_raskid' stavljena je oznaka 'Y' ako je 'datum_zatvaranja' + 10 $<$ 'planirani datum_zatvaranja'. Zatim smo primijetili da je u nekim izvještajima starost veća od 900 ili negativna pa smo u takvim slučajevima postavili vrijednost na prosjek preostalih vrijednosti klijenata istog tipa.

Nadalje, pozabavili smo se vrijednostima koje nedostaju. Primijetili smo da se visine kamata razlikuju ovisno o tipu kamate, vrsti klijenta i vrsti proizvoda, pa smo izračunali prosjek visina kamata kako bismo nadopunili nepoznate vrijednosti u tom stupcu. Kako bismo popunili stupce 'planirani_datum_zatvaranja' i 'datum_zatvaranja', prošli smo kroz sve kvartalne izvještaje za određenu partiju te smo stavili pronađenu vrijednost ako takva postoji. Da bismo datume prikazali kao numeričke podatke, zamijenjeni su vremenom proteklim od 1.1.1970. (u sekundama).

Primijetili smo da je ciljna varijabla jednaka 'N' do zadnjih izvještaja, bez obzira na to je li ugovor u konačnici prijevremeno raskinut ili ne, pa smo taj stupac ispravili postavljanjem na vrijednosti u zadnjem izvještaju.

Definirali smo dvije nove značajke: ’produljenja’ te ’planirana_duljina_ugovora’. Naime, za svako produljenje ugovora, izvještaji imaju novi datum otvaranja. Svako produljenje ugovora (osim eventualno zadnjeg) je označeno kao pravovremeno raskinuti ugovor. Takvi ugovori se pojavljuju u Mozgalovom skupu za testiranje pa smo odlučili razmatrati svako produljenje posebno, a ne sve skupa kao jednu instancu. Značajku ’planirana_duljina_ugovora’ definirali smo kao razliku između vrijednosti u stupcima ’planirani_datum_zatvaranja’ i ’datum_otvaranja’ (u sekundama).

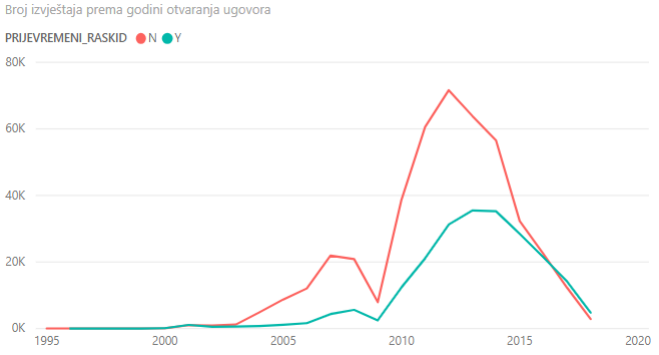
Izbrisani su podaci o ugovorima koji još nisu zatvoreni te kojima nedostaje planirani datum zatvaranja, kao i stupci koji se odnose na same izvještaje (datum izvještavanja, stanje na kraju kvartala i stanje na kraju prethodnog kvartala). Nakon svih transformacija i definicija ostali su duplikati pa smo ostatak skupa podataka reducirali na jedan redak po ugovoru za dani id klijenta, oznaku partije i datum otvaranja.

### *C. Vizualizacija podataka*

Za vizualizaciju značajki, napravili smo kopiju podataka i izbacili značajke vezane uz izvještaje kako bismo saželi skup podataka (’datum_izvjestavanja’, ’datum_zatvaranja’, ’planirani_datum_zatvaranja’, ’stanje_na_kraju_kvartala’, ’stanje_na_kraju_prethodnog_kvartala’ ). Time smo dobili oko 663000 izvještaja pomoću kojih smo, koristeći PowerBI te Orange, došli do nekih generalnih zaključaka o skupu podataka.

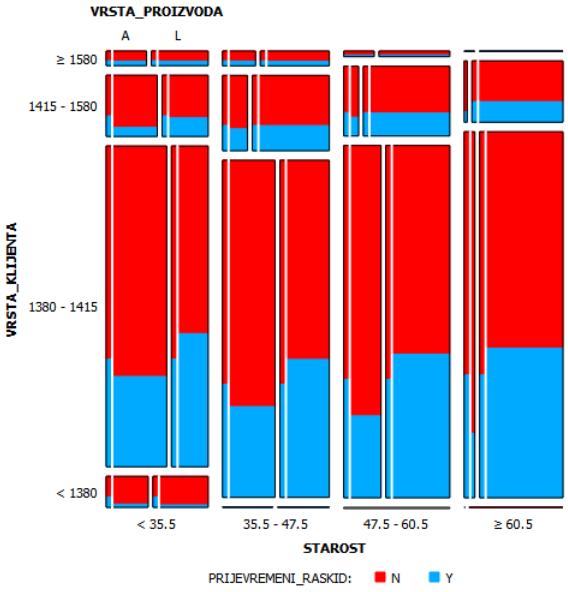
Nadalje, pomoću filtera na kategoriju ’prijevremeni_raskid’ u Power BI-u na kojem je moguće odabrati kategorije ’Y’, ’N’ ili ’All’, separirali smo vizualizacije napravljene za dobiveni skup ovisno o tome je li ugovor prijevremeno raskinut ili ne.

Kako bismo dobili bolji uvid s čime zapravo baratamo, odlučili smo prikazati broj izvještaja prema godini otvaranja ugovora uz podjelu ovisno o prijevremenom raskidu.



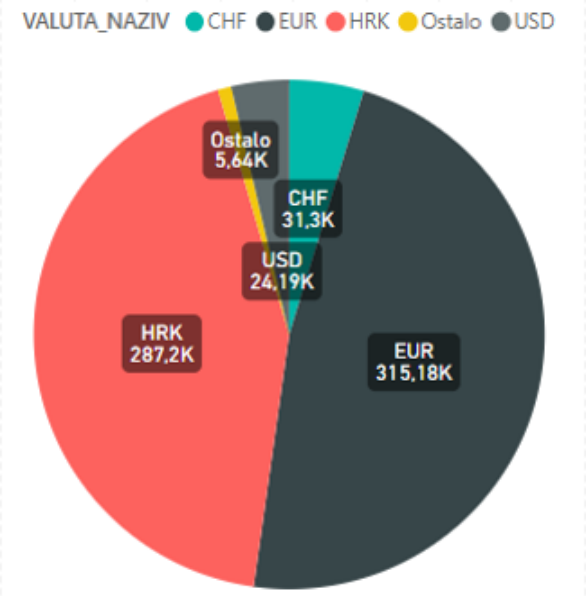

Ovaj graf nam je bio zanimljiv jer se na njemu može vidjeti socio-ekonomsko stanje u državi u zadnja 3 desetljeća. Najmanje ugovora otvoreno je u ratnom vremenu, nakon čega dolazi do rasta broja otvorenih ugovora sve do ekonomske krize 2009.god. Nakon toga uočava se ponovni rast sve do 2015. što je moguća posljedica iseljavanja stanovništva. Iz ovog grafa zaključili smo da će makroekonomski pokazatelji biti bitni za daljnju analizu.

Sljedeće, ovaj mozaik nam je bio veoma pregledan jer je omogućio kategorizaciju raskida. Klijenti su grupirani prema vrsti i starosti te dodatno prema vrsti proizvoda kojeg su odabrali (štednja ili kredit). Konačno, bojama se za svaku kategoriju vidi omjer prijevremeno i pravovremeno raskinutih ugovora.

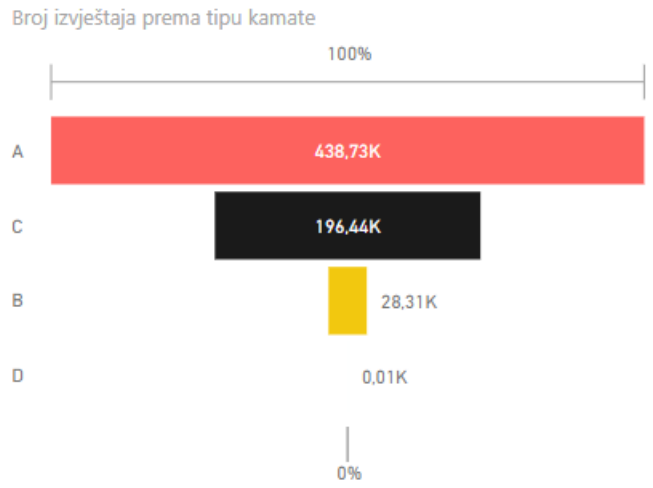



Najviše je klijenata 1380-1415, što može sugerirati fizičke osobe. Što se tiče proizvoda, za svaku kategoriju osim za mlađe od 25 više je vrste L što označava štednju. Dakle, mlađe i pravne osobe češće biraju podizati kredit. Slično, za svaku vrstu klijenata i stariju životnu dob, znatno je manje proizvoda A što sugerira da osobe starije životne dobi vjerojatnije biraju štednju.

Dodatno, zanimala nas je distribucija određenih klasa u skupu. Na pie chartu vidimo da su najzastupljenije valute bile EUR i HRK.



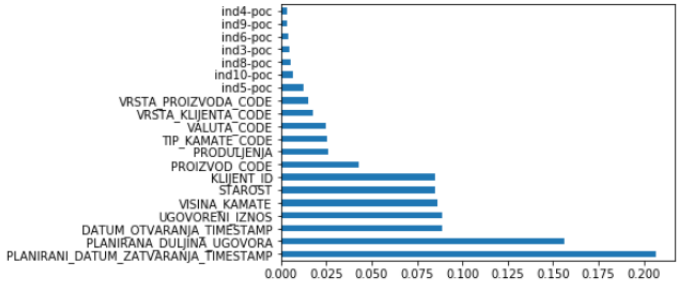
Kada promotrimo broj izvještaja prema tipu kamate, uočava se nebalansiranost jer je daleko najviše kamate tipa A, dok se kamata tipa D pojavljuje tek u malom broju izvještaja.

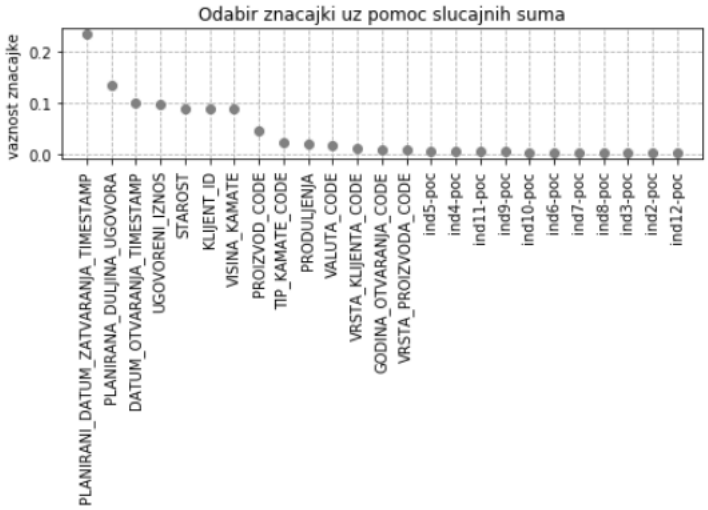



## III. Izvlačenje značajki

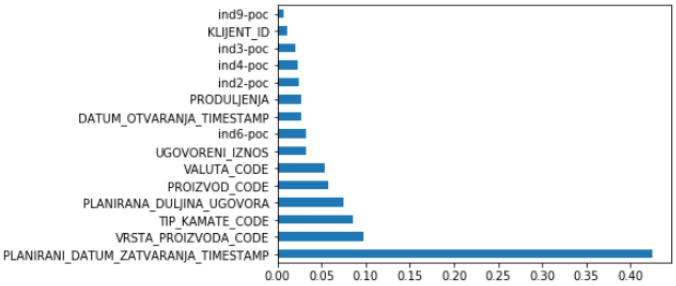
Iskoristili smo makroekonomske pokazatelje sa stranice HNB-a, RBA i DZS-a te tako dobili 12 dodatnih značajki na godišnjoj razini: BDP po stanovniku (u EUR), BDP- realna godišnja stopa promjene (u %), prosječna godišnja stopa inflacije potrošačkih cijena, tekući račun platne bilance (u % BDP-a), uvoz/izvoz robe i usluga (u % BDP-a), inozemni dug (u % BDPa), prosječni devizni tečaj (HRK: 1 EUR), prosječni devizni tečaj (HRK: 1 USD), neto pozajmljivanje (+)/zaduživanje (-) konsolidirane opće države (u % BDP-a), dug opće države (u % BDP-a) te stopa zaposlenosti (prema definiciji ILO-a, stanovništvo starije od 15 godina). Za jednostavniji unos pri evaluaciji značajki, promijenili smo imena redom u ind1-poc – ind12-poc. Makroekonomski pokazatelji su dani po godinama pa smo samo spajanje napravile na temelju godina otvaranja ugovora, koje smo zasebno definirali kao značajku.

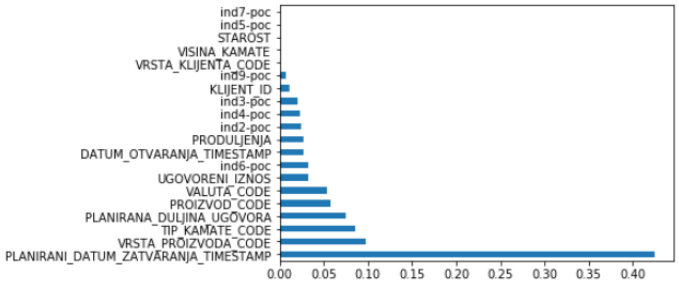
Kako bismo odredili koje ćemo značajke koristiti za treniranje modela, prvo smo iskoristili univarijatni odabir pomoću $\chi^2$ statistike i slučajnih šuma. Dobiveni su sljedeći rezultati:




Kao bitna značajka osobito se ističu 'planirani_datum_zatvaranja' i 'planirana_duljina_ugovora'. Osim toga, iskoristili smo i XGBClassifier i ponovno rangirali 15 i 20 značajki po važnosti:



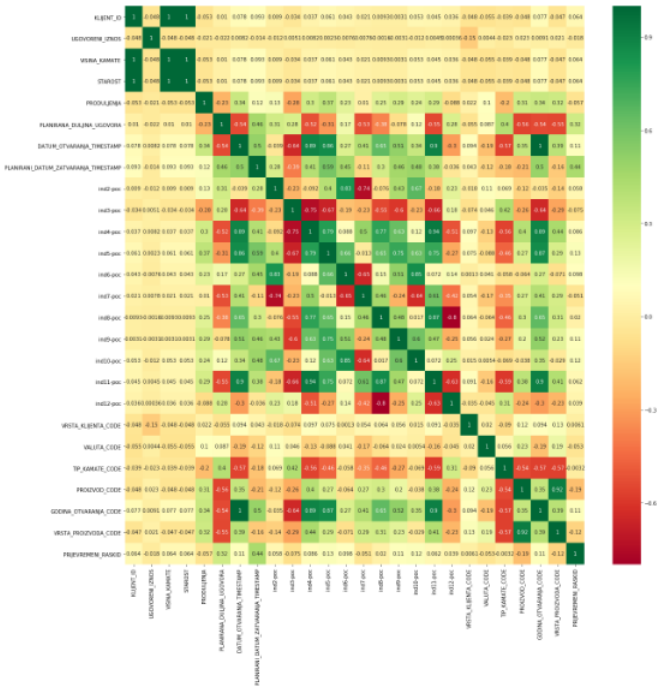

U ovom slučaju, kao najznačajnija značajka se ističe 'planirani_datum_zatvaranja_timestamp' ('_code' označava enkodirane kategoričke varijable.)

Konačno, isprobali smo i rekurzivnu eliminaciju značajki uz cross validaciju te smo rezultat te metode odabrali kao naš konačan odabir značajki, njih ukupno 17 od originalnih 25.

- KLIJENT_ID
- UGOVORENI_IZNOS
- VISINA_KAMATE
- STAROST
- PRODULJENJA
- PLANIRANA_DULJINA_ZATVARANJA_TIMESTAMP

- ind4-poc (tekući račun platne bilance u postotku BDP-a)
- ind5-poc (uvoz/izvoz robe i usluga)
- ind11-poc (neto pozajmljivanje/zaduživanje države u postotku BDP-a)
- VRSTA_KLIJENTA_CODE
- VALUTA_CODE
- TIP_KAMATE_CODE
- PROIZVOD_CODE
- GODINA_OTVARANJA_CODE
- VRSTA_PROIZVODA_CODE

Također, promatrali smo koreliranost značajki koristeći heatmap. Koristili smo svih 25 značajki zajedno sa ciljnom značajkom i promatrali koreliranost za svaki par značajki.



Kao što smo i očekivali, postoji velika pozitivna koreliranost između određenih makroekonomskih značajki, npr. 0,93 između značajki inozemni dug i tekući račun platne bilance. S druge strane, očekivano je i da su neke negativno korelirane, npr. -0,84 između prosječnog deviznog tečaja i stope zaposlenosti. Neobičan rezultat je koreliranost 1 između značajki 'starost', 'klijent_id' te 'visina_kamate'. Visoka koreliranost djeluje nepovoljno na model pa neke značajke neće biti uvrštene kasnije u model.

## IV. Opis metode i pristupa za rješavanje problema

Zbog nebalansiranosti skupa podataka i promatranja rješenja problema sličnih našem, odlučili smo koristiti Random Forest Classifier, XGBoost Classifier te neuronske mreže. U rješavanje problema krenuli smo algoritmom Random Forest Classifier, ansambl algoritmom za klasifikaciju. Prilikom treniranja, Random Forest algoritam stvara velik broj stabala, od kojih se svako trenira na određenom broju uzoraka trening seta i vrši pretragu po slučajno generiranom podskupu ulaznih varijabli kako bi odredio mjesto na kojem će se razgranati. Za klasifikaciju svako stablo daje glas jednoj od klasa. Izlaz klasifikatora ovisi o broju glasova stabala svakoj pojedinoj klasi. Mijenjanjem broja stabala(n_estimators), pokušavali smo dobiti što veću preciznost i F1-score.

Dalje smo nastavili s algoritmom XGBoost (eXtreme Gradient Boosting), koji je izrazito popularan algoritam za klasifikaciju, regresiju i ranking probleme. XGBoost je implementacija gradient boosting (GB) strojeva i ima podršku za tri forme gradient boosting-a (Gradient Boosting (GB), Stochastic GB i Regularized GB). Algoritam stvara predikcijski model u obliku ansambla slabih predikcijskih problema, to jest stabala odlučivanja. Za učenje smo podešavali sljedeće parametre: learning_rate (stopa učenja), n_estimators(broj stabala), max_depth(maksimalna dubina stabla), min_child_weight (najmanja suma težina u čvoru koji je dijete), gamma (najmanji loss potreban za grananje čvora), subsample (omjer subsamplinga), colsample_bytree (omjer subsampling-a stupaca za svako stablo), objective (specifikacija problema i objekta učenja), nthread (broj paralelnih dretvi), scale_pos_weight (kontrolira balans pozitivnih i negativnih težina), seed (početna točka pri generiranju slučajnih nizova brojeva).

Zadnja metoda koju smo koristili su neuronske mreže, međusobno povezani umjetni neuroni. Neuronska mreža može se opisati kao umjetna preslika ljudskog mozga kojom se nastoji simulirati postupak učenja. Moć analize pohranjena je u snazi veza između pojedinih neurona, tj. težinama do kojih se dolazi učenjem iz skupa podataka. Mreže se sastoje od ulaznog sloja, skrivenih slojeva i izlaznog sloja. Trening skup učili smo na mrežama s tri ili četiri sloja (jedan ili dva skrivena), različitim brojem neurona u ulaznom i skrivenom sloju i jednim neuronom u izlaznom sloju, različitim funkcijama aktivacije na ulaznim i skrivenim slojevima, funkcijom aktivacije 'sigmoid' na izlaznom sloju te različitim brojem epoha učenja i različitim veličinama parametra batch_size.

Kao što smo očekivali, najbolje rezultate dobili smo koristeći algoritam XGBoost, kojim se u posljednje vrijeme postižu jako dobri rezultat i pobjeđuju natjecanja u području data science i ML kao što je Kaggle.

## V. Prikaz rezultata

Dodatno smo odlučili usporediti sedam metoda train-test splitom: Random Forest, Extra Trees, logističku regresiju, Decision Tree, AdaBoost, Gradient Boosting i Bagging classifier. Usporedili smo ih na skupu za treniranje sa i bez odabira značajki, te s raznim tehnikama sampliranja i bez ikakvog sampliranja.

Kao što je već navedeno, skup podataka je dosta nebalansiran; nakon čišćenja podataka imamo omjer 1:4 u korist vrijednosti ciljne varijable 'Y'. Od tehnika sampliranja isprobano je slučajni oversampling i undersampling, SMOTE i

ADASYN oversampling. Sampling tehnike su provedene samo na skupovima za treniranje unutar cross-validacije. SMOTE i ADASYN su malo naprednije metode za sampliranje; generiraju nove uzorke iz manjinske klase na temelju k-nn algoritma. Za sampliranje nam je bila potrebna biblioteka imbalanced-learn.

Najboljim se pokazalo treniranje na skupu s odabranim značajkama i bez ikakvog sampliranja:

| MLA Name | MLA Train Accuracy | MLA Test Accuracy | MLA F1-score | MLA AUC |
|---|---|---|---|---|
| RandomForestClassifier | 0.9861 | 0.8806 | 0.576048 | 0.714293 |
| BaggingClassifier | 0.9864 | 0.8802 | 0.581873 | 0.719063 |
| GradientBoostingClassifier | 0.8798 | 0.8794 | 0.519971 | 0.678740 |
| ExtraTreesClassifier | 0.9999 | 0.8722 | 0.554537 | 0.705742 |
| AdaBoostClassifier | 0.8717 | 0.8721 | 0.476437 | 0.658604 |
| LogisticRegressionCV | 0.8669 | 0.8668 | 0.458583 | 0.651473 |
| DecisionTreeClassifier | 0.9999 | 0.8318 | 0.536207 | 0.720147 |

Slika 1. Rezultati evaluacije raznih modela train-test split metodom

Tehnike oversampling-a su imale par postotaka nižu točnost i f1-score, dok je random undersampling imao točnost lošiju za oko 10%, a f1-score lošiji za par posto.

Nakon toga smo odlučili detaljni istražiti modele Random Forest i XGBoost, uz dodatno pretraživanje parametara. Najbolji postignuti rezultati su prikazani u sljedećoj tablici:

| | Accuracy | f1-score |
|---|---|---|
| Random Forest | 0.88 | 0.58 |
| XGBoost | 0.88 | 0.59 |

Tablica I
REZULTATI PRETRAŽIVANJA PARAMETARA ZA RF I XGB

Na kraju smo odlučili isprobati neuronske mreže koje su se bez posebnog dodatnog podešavanja pokazale kao dobar model (točnost od 87%, a f1-score 55%). Jedino je isprobano postavljanje težina na klase i uz to je f1-score poboljšan za 5%. Zbog manjka vremena nismo se detaljnije bavili neuronskim mrežama, ali smatramo da bi se mogao postići puno bolji rezultat.

Organizatori Mozgala su u raznim fazama dali skup za treniranje, testiranje i validaciju. Oznake skupova za testiranje i validaciju nam nisu bile dostupne, već se evaluacija modela odvijala online, na web-servisu.

Za potrebe tog testiranja smo se fokusirali na tri metode: random forest, XGBoost i neuronske mreže. Random forest je imao točnost 63%, a f1-score 61%, neuronske mreže točnost 66%, f1-score 64%, dok se XGBoost na tom skupu pokazao znatno boljim - pretraživanjem parametara smo postigli točnost 75%, a f1-score 78%. Naime, skup za testiranje na web-servisu je imao drukčiju distribuciju klasa nego skup za treniranje ('Y' i 'N' u omjeru otprilike 3:2), dok je originalni skup za treniranje bio više nebalansiran pa tako i skupovi za treniranje i testiranje dobiveni iz tog skupa.

Rezultat na skupu za validaciju je bio znatno lošiji (točnost 58%, f1-score 52%) gdje smo imali samo dva pokušaja za evaluaciju. Tu smatramo da je bila naša greška u tome što smo konstatno pokušavali unaprijediti sam skup za treniranje (u smislu čišćenja i vremenskih serija) tijekom faze treniranja pa nam je model u nekom trenutku bio lošiji od prethodnih te se to odrazilo i na fazu validacije.

| | Accuracy | f1-score |
|---|---|---|
| model testing | 0.75 | 0.78 |
| model validation | 0.58 | 0.52 |

Tablica II
REZULTATI NA NATJECANJU

## VI. OSVRT NA DRUGE PRISTUPE

Ovaj problem bio je u sklopu natjecanja Mozgalo, na kojem je tim u finalu osvojio šesto mjesto. Naše rješenje koje se vrednovalo na natjecanju bio je XGBoost algoritam, koji je koristilo mnogo timova. Sličan pristup imalo je još timova koji su koristili CatBoost i undersampling. Jako zanimljiv i inovativan je bio pristup drugoplasiranog tima koji je numeričke varijable enkodirao pomoću percentilnih rangova te poslije toga koristi regresijski ansambl s LGBM klasifikatorom. Također je zanimljivo rješenje pobjedničkog tima koji je odbacio 40% dataseta te koristio dva stacking sloja: u prvom je bilo korišteno pet lineranih, dvije k-NN, tri tree-based metode te dvije neuronske mreže, dok su u drugom sloju bile korištene po jedna linearna i tree-based metoda te jedna neuronska mreža.

Ono čime se naš tim na natjecanju Mozgalo istaknuo bio je različit odabir integriranja makroekonomskih značajki, dodavanje nekih novih značajki te jako informativne vizualizacije.

## VII. MOGUĆI BUDUĆI NASTAVAK ISTRAŽIVANJA

Prilikom budućeg rješavanja problema, mogli bismo još isprobati metodu privilegiranog učenja koju je osmislio Vapnik (SVM+) kako bi se moglo iskoristiti podatke o kvartalima, a koji su ovdje zanemareni. Od ostalih metoda mogli bismo još razmotriti neuronske mreže za vremenske serije (LSTM) te na koji način bismo ih mogli iskoristiti na podacima o kvartalima.

Osim toga, mogli bismo iskoristiti podatke o makroekonomskim podacima kroz cijelo razdoblje ugovora, ne samo na početku kao što je u ovom rješenju napravljeno.

Također, koristili smo makroekonomske podatke samo za Hrvatsku (osim tečaja valuta), a mogli bismo razmotriti i globalne makroekonomske indikatore jer svjetsko tržište ima veliki utjecaj na ekonomiju naše zemlje.

## LITERATURA


[1] [Online]. Available: https://www.dzs.hr/
[2] [Online]. Available: https://www.rba.hr/mala-poduzeca-i-obrtnici/istrazivanja-i-analize/publikacije
[3] [Online]. Available: https://www.hnb.hr/statistika/glavni-makroekonomski-indikatori
[4] J. H. Friedman, R. Tibshirani, and T. Hastie, *Elements of Statistical Learning*, 2008.
[5] J. D. Hamilton, *Time Series Analysis*.

[6] T. Pavić Kramarić, *Osnove financija*.
[7] O. Blanchard, *Macroeconomics*, 3rd ed., 2015.
[8] R. Thomas, *Modern econometrics*.
[9] D. Findlay and O. Blanchard, *Macroeconomics-Study Guide*.
[10] [Online]. Available: https://www.coursera.org/learn/machine-learning
[11] [Online]. Available: http://stakana.com/
[12] N. Derby and M. Keintz, "Reducing customer attrition with predictive analytics for financial institutions," *MWSUG 2016 - Paper BI04*, 2016.
[13] L. Baldassini and J. A. Rodriguez Serrano, "client2vec: Towards systematic baselines for banking applications," 2018.
[14] M. Galovič, "Time series forecasting with rnn's," 2018.
[15] V. Avon, "Machine learning techniques for customer churn prediction in banking environments," Master's thesis, Universit'a degli Studi di Padova, 2016.