1. Opis odabranog pristupa za rješavanje problema

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.

Prvi korak u rješavanju ovog problema bila je podrobna analiza podataka. Budući da su podatci većinom pod šiframa, eksplorativnom analizom bilo je potrebno izvesti zaključke o vrsti klijenata i njihovim ugovorima s bankom. Jedan od problema s kojim smo se susreli je činjenica da je skup podataka nebalansiran i nepotpun, tj. neke od komponenti imaju znatan broj nepostojećih i/ili nelogičnih unosa (starost, visina_kamate, datum_zatvaranja, itd.). Unošenjem dodatnih makroekonomskih parametara, transformacijom podataka te koristeći razne metode vizualizacije, bilo je moguće detektirati koji od klijenata su zaista fizičke osobe, a koji pravne.

Drugi problem je to što skup za treniranje sadrži više značajki nego skup za testiranje. U ovom rješenju te značajke nisu korištene. Potrebno je bilo detaljno i pažljivo pripremiti podatke, jer se neke (potpune) informacije o ugovoru ne nalaze u svim izvještajima, već najčešće u zadnjem (kao oznaka prijevremenog raskida i datum zatvaranja).

Nakon pripreme podataka, enkodiranjem smo pripremili kategoričke varijable (vrsta_klijenta, valuta, tip_kamate i sl.) za treniranje modela. Kao posljednji korak u pripremi, provedena je i min-max scaling metoda. Kako bismo dobili što veću uspješnost modela i smanjili overfitting, napravili smo feature selection. Metode koje smo odabrali su univarijatni odabir značajki te matricu korelacije s heatmapom. Zbog velikog broja varijabli, u univarijatnom odabiru isprobali smo odabir 20 i 15 najvažnijih značajki. Što se tiče heatmape, primijetili smo veliku koreliranost između određenih značajki, primjerice 'godina_otvaranja' i vanjski dug postižu vrijednost od 0,89.

Konačno, budući da se radi o klasifikacijskom problemu, za izradu modela koristili smo XGBoost. XGBoost nam je bio primarni izbor jer smanjuje varijabilnost i reducira pristranost. Dodatno, razmotrili smo slučajne šume te neuronske mreže, no najuspješniji rezultati postignuti su koristeći XGBoost uz hyperparameter tuning.

2. Opis dataseta

a. Dataset

Na raspolaganju su nam podatci o transakcijskim računima pravnih osoba i/ili stambenih kredita fizičkih osoba u nekoliko vremenskih nizova. Za svaku fizičku/pravnu osobu dostupno je 17 parametara, kao što su ugovoreni iznos, datum otvaranja, visina kamate i valuta. Problem je klasifikacijski: ciljna značajka je kategorička varijabla koja govori je li ugovor (kredit ili štednja) prijevremeno raskinut

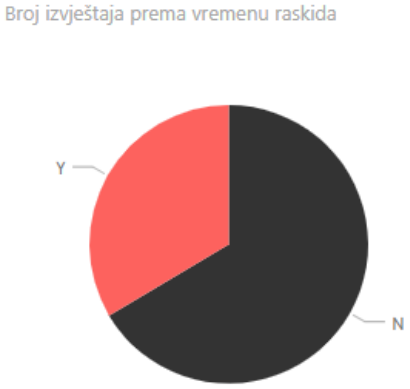
ili ne. U našem skupu za treniranje otprilike 70% ugovora je pravovremeno raskinuto, dok je 30% raskinuto prijevremeno.

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.

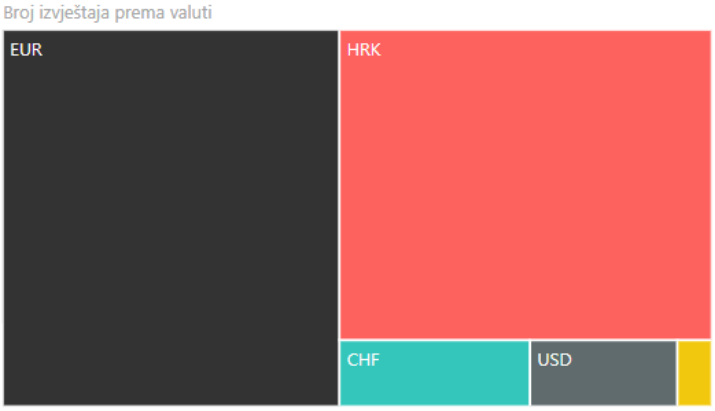
Nadalje, pozabavili smo se vrijednostima koje nedostaju. Primijetili smo da se visine kamata razlikuju ovisno o tipu kamate, vrsti klijenta i vrsti proizvoda, pa smo izračunali prosjek visina kamata kako bismo nadopunili nepoznate vrijednosti u tom stupcu. Kako bi popunili stupac 'planirani_datum_zatvaranja', prošli smo kroz sve kvartalne izvještaje za određenu partiju te smo stavili pronađenu vrijednost ako takva postoji.

b. Vizualizacija

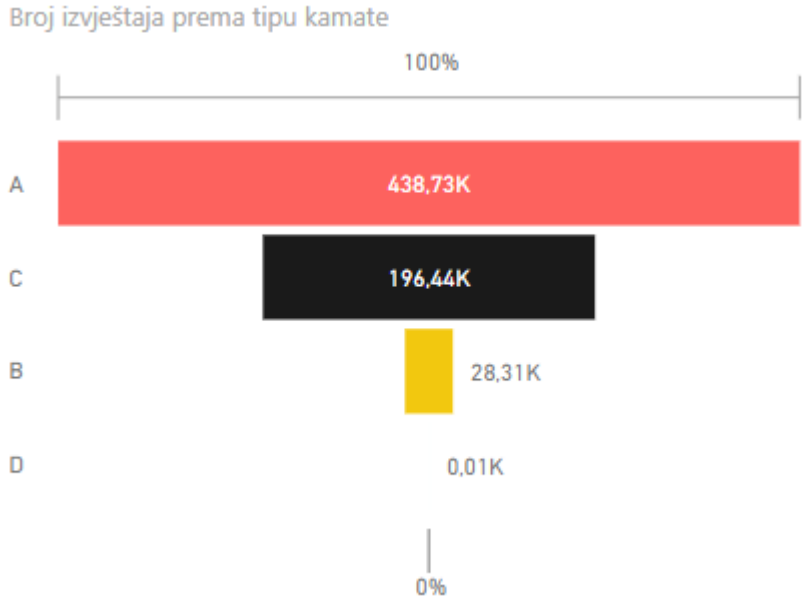
Za vizualizaciju značajki, napravili smo kopiju podataka i izbacili značajke vezane uz izvještaje kako bismo saželi skup podataka ('datum_izvjestavanja', 'datum_zatvaranja', 'planirani_datum_zatvaranja', 'stanje_na_kraju_kvartala', 'stanje_na_kraju_prethodnog_kvartala'). Nadalje, separirali smo dobiveni skup ovisno o tome je li ugovor prijevremeno raskinut ili ne. Koristeći PowerBI te Orange, došli smo do nekih generalnih zaključaka o skupu podataka.



Iz sljedećeg pie chart-a, jasno je da je skup podataka nebalansiran budući da ima znatno više ugovora koji nisu prijevremeno raskinuti (oko 70%).

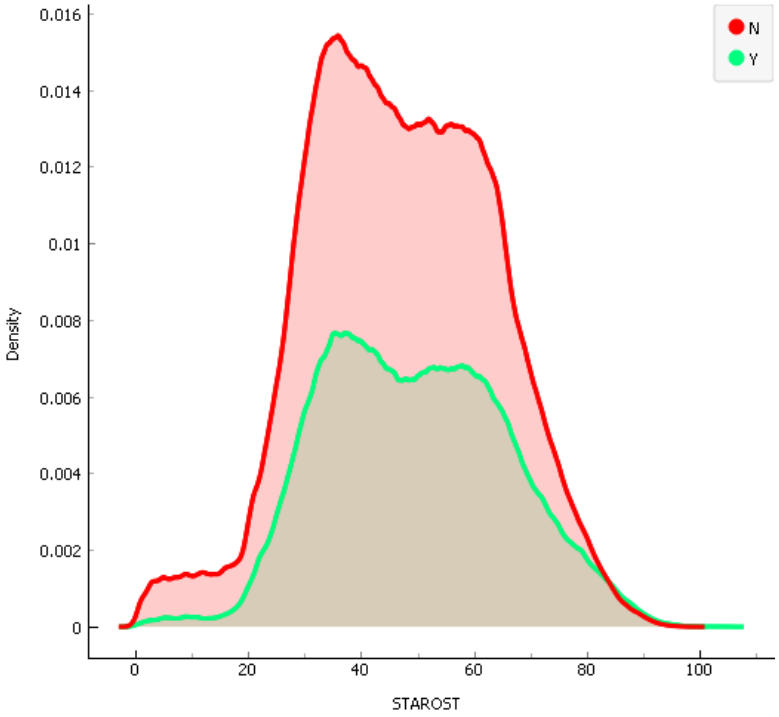


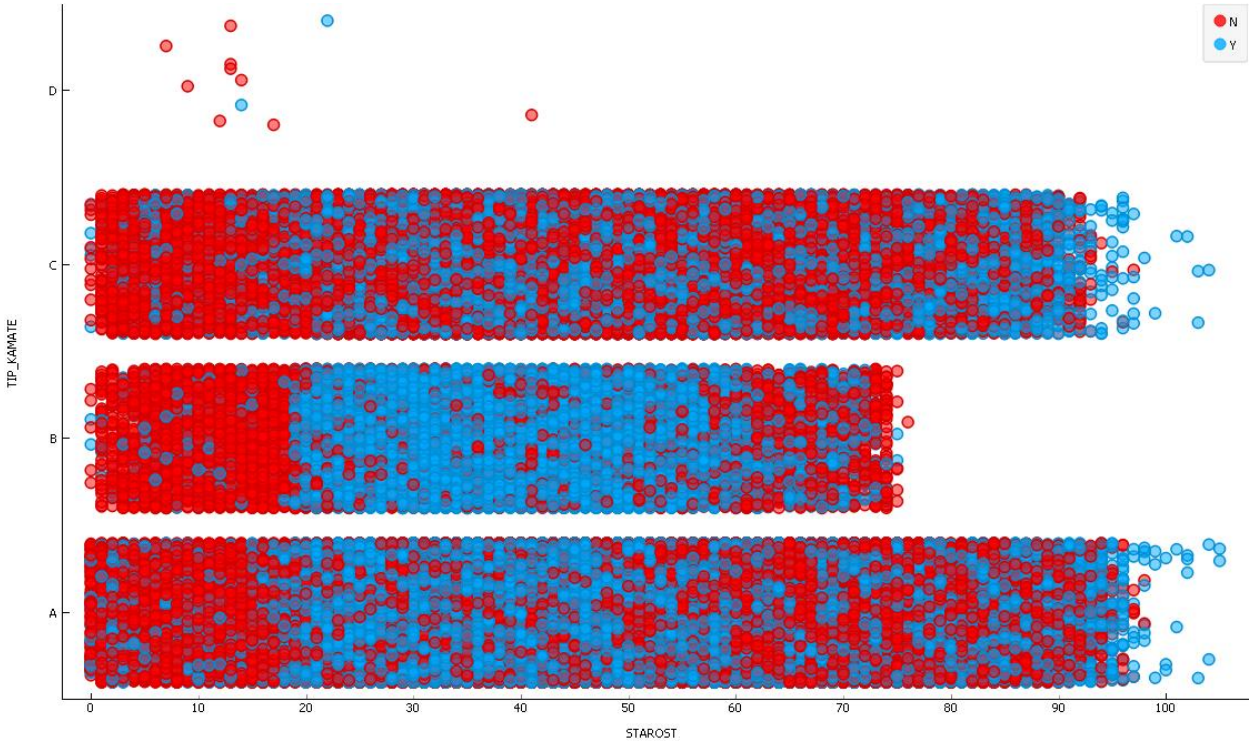
Nadalje, ako promotrimo zastupljenost valuta u izvještajima, prevladavaju EUR i HRK, dok su USD i CHF prisutne u znatno manjoj mjeri.
Sljedeće, promatrali smo zastupljenost tipova kamate.



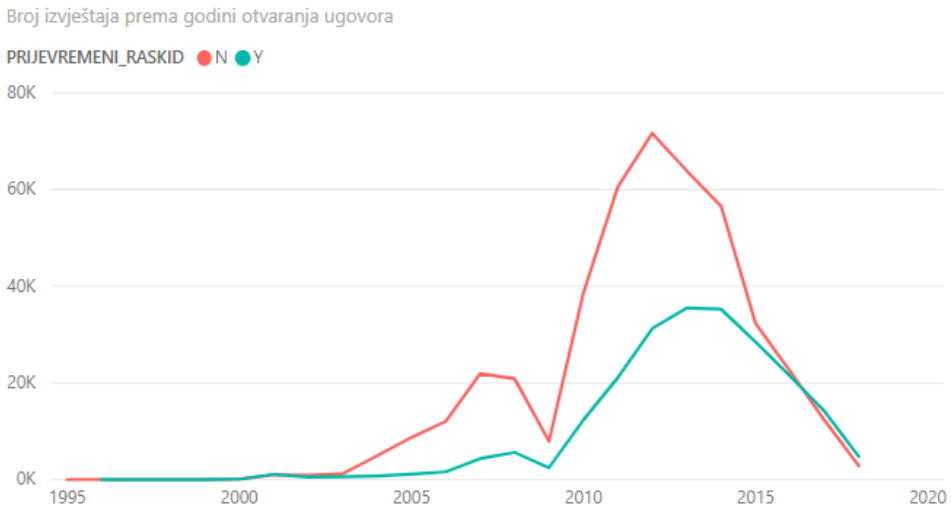
Najzastupljeniji je tip kamate A, dok je zanimljivo bilo za uočiti da je tip kamate D prisutan u neznatnoj mjeri u usporedbi s ukupnim brojem ugovora. Nakon toga, analizirali smo klijente banke.

Na sljedećem grafu prikazana je distribucija starosti separiranog skupa klijenata. Grafovi podsjećaju na graf normalne distribucije i može se uočiti da je najmanje klijenata mlađih od 20 i starijih od 80 godina, što je i očekivano. U globalu, za sve dobne skupine više ugovora ne bude prijevremeno raskinuto. Moguće je pretpostaviti da su klijenti mlađi od 20 godina pravne osobe.

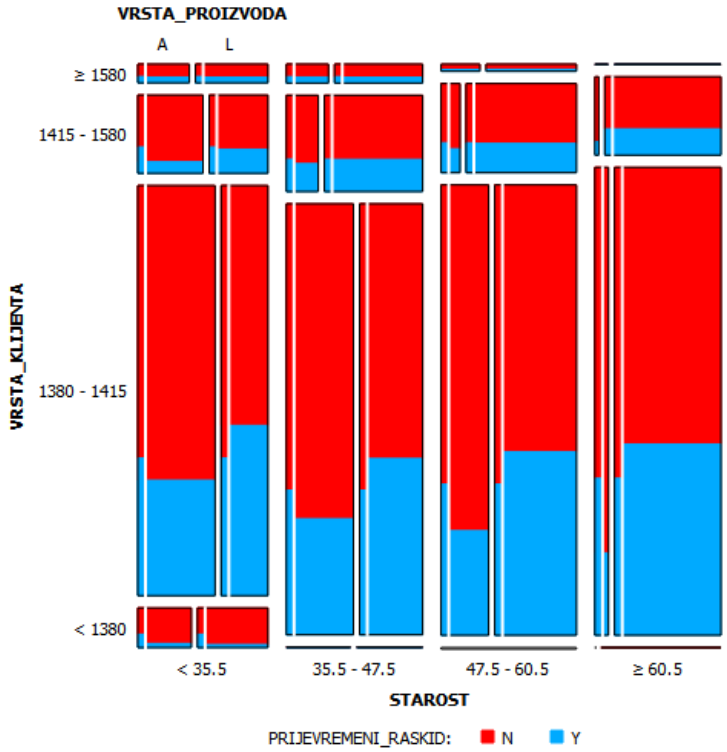




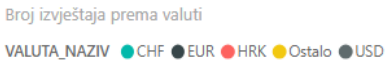
Ako u daljnju analizu dodamo i tip kamate, dobiva se ovakav scatter plot. Za svaki tip kamate, tendencija je klijenata mlađih od 20 godina (prema našoj pretpostavki većinom pravnih osoba) da se ugovor ne raskida prijevremeno. Klijenti srednje životne dobi najviše ranije raskidaju ugovore, što se osobito ističe za tip kamate 'B'. Odlučili smo i promotriti datume otvaranja ugovora.

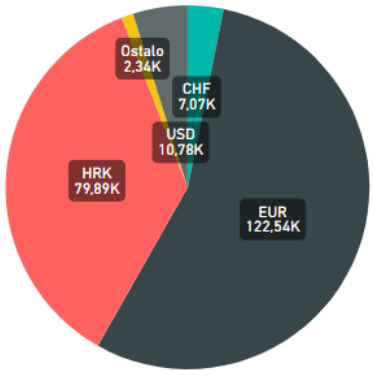


Ovaj graf nam je bio zanimljiv jer se u njemu odražava i socio-ekonomsko stanje u državi u zadnja tri desetljeća. Najmanji broj ugovora je u ratnom vremenu i par godina nakon, što je i logičan slijed događaja. Nakon toga dolazi do rasta u broju ugovora sve do značajnog pada oko 2009. godine, tj.u vrijeme ekonomske krize. Broj ugovora tada bilježi rast sve do postupnog opadanja od 2015. godine (kada broj prijevremeno raskinutih ugovora čak i premaši neraskinute), što bi mogla biti posljedica iseljavanja stanovništva. Dakle, makroekonomski pokazatelji će nam biti bitni za model. Konačno, particionirali smo vrstu proizvoda, vrstu klijenta te starost u odnosu na prijevremeni raskid. Najviše je vrste klijenata 1380 – 1415, što bi moglo sugerirati da se radi o fizičkim osobama. Što se tiče vrste proizvoda, za svaku kategoriju (osim za mlađe od 35.5 godina) više je vrste 'L' što označava štednju. Može se zaključiti da će osobe mlađe životne dobi (kao i pravne osobe) vjerojatnije podizati kredite. Također, za svaku vrstu klijenta i za stariju životnu dob znatno je manje vrste proizvoda 'A'. Može se zaključiti da će osobe starije životne dobi vjerojatnije birati štednju.

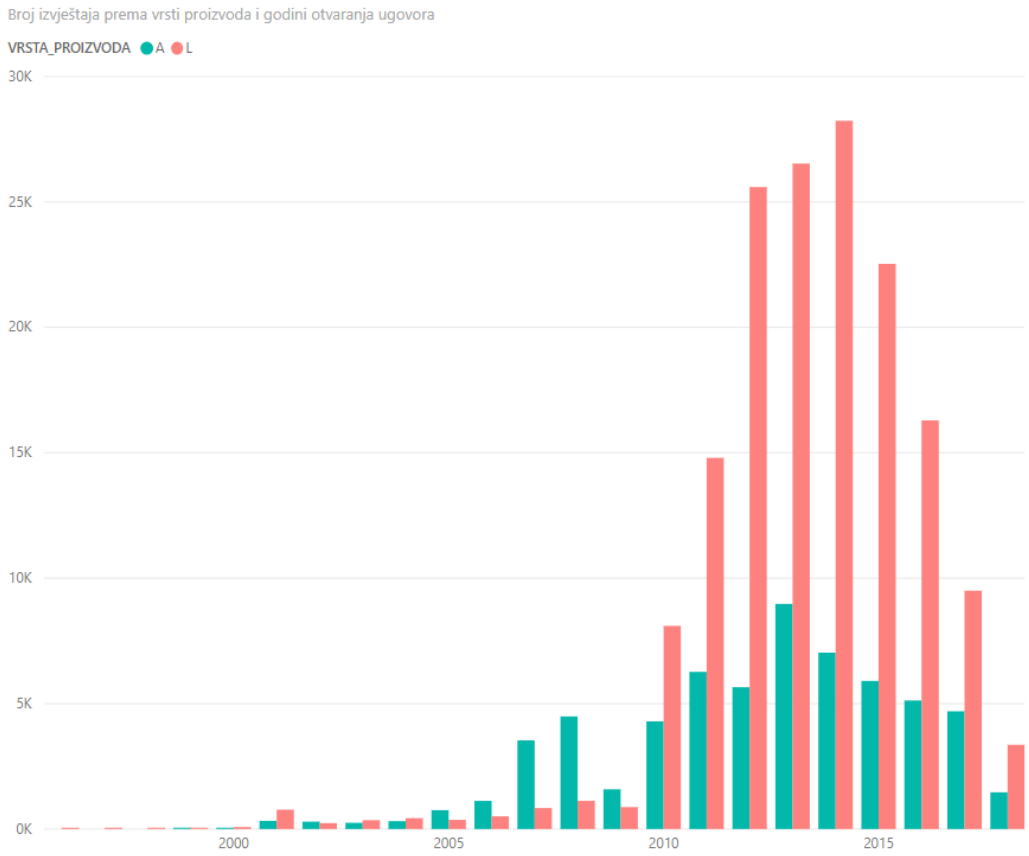


Konačno, u PowerBI napravili smo zasebnu obradu za ugovore koji su prijevremeno raskinuti i za one koji nisu. Kao primjer navodimo grafove prve kategorije.





Ovaj pie chart prikazuje učestalost vrsta valuta u prijevremeno raskinutim ugovorima. Kao i u analizi cjelokupnog skupa podataka, prevladavaju EUR i HRK.



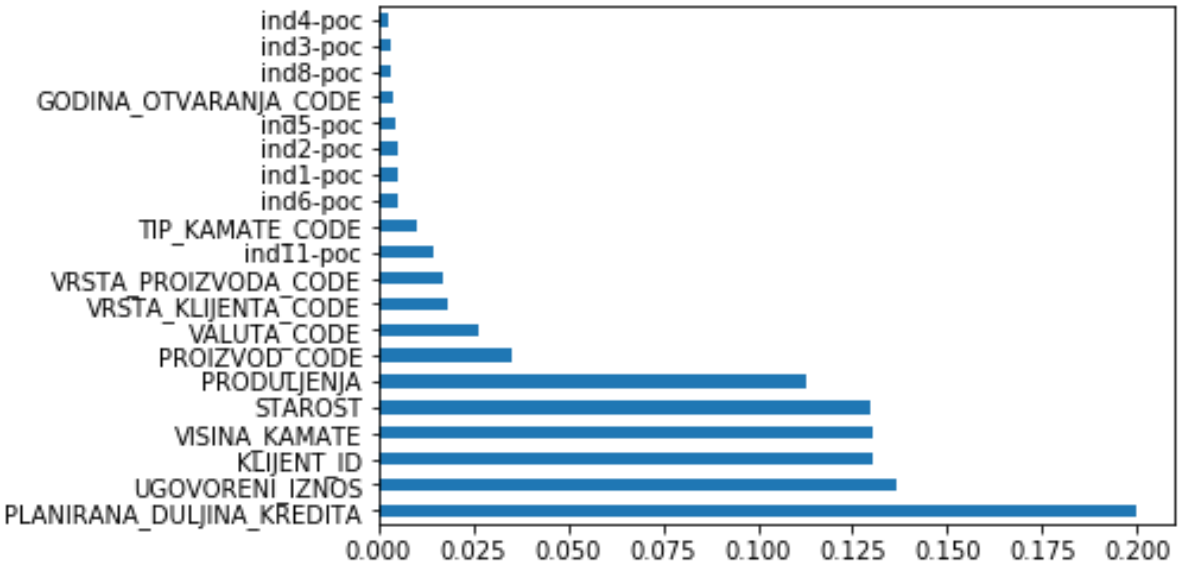
Dodatno, analizirali smo vrste proizvoda kroz vrijeme otvaranja ugovora. Prikazani podatci slijede trend prethodno opisanog socio-ekonomskog stanja. Zanimljivo je za primijetiti razliku u odabiru vrste proizvoda prije i poslije 2010.godine.

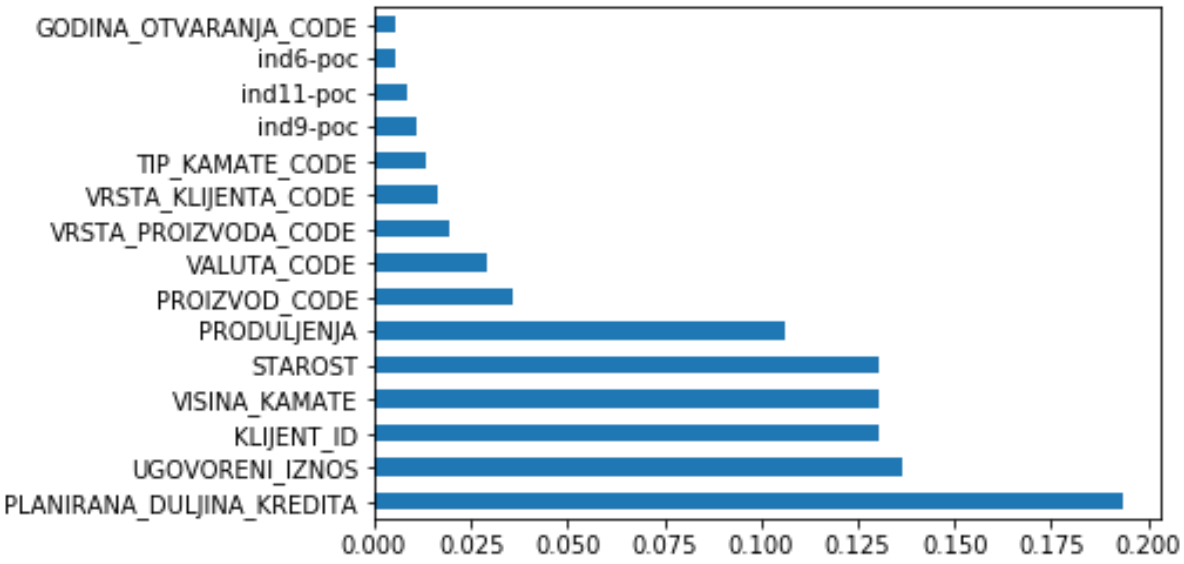
3. Izvlačenje značajki i njihova značajnost u korištenom modelu

Iskoristili smo makroekonomske pokazatelje sa stranice HNB-a 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 % BDP-a), 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. Nadalje, prvotne dobivene značajke transformirali smo u nove.

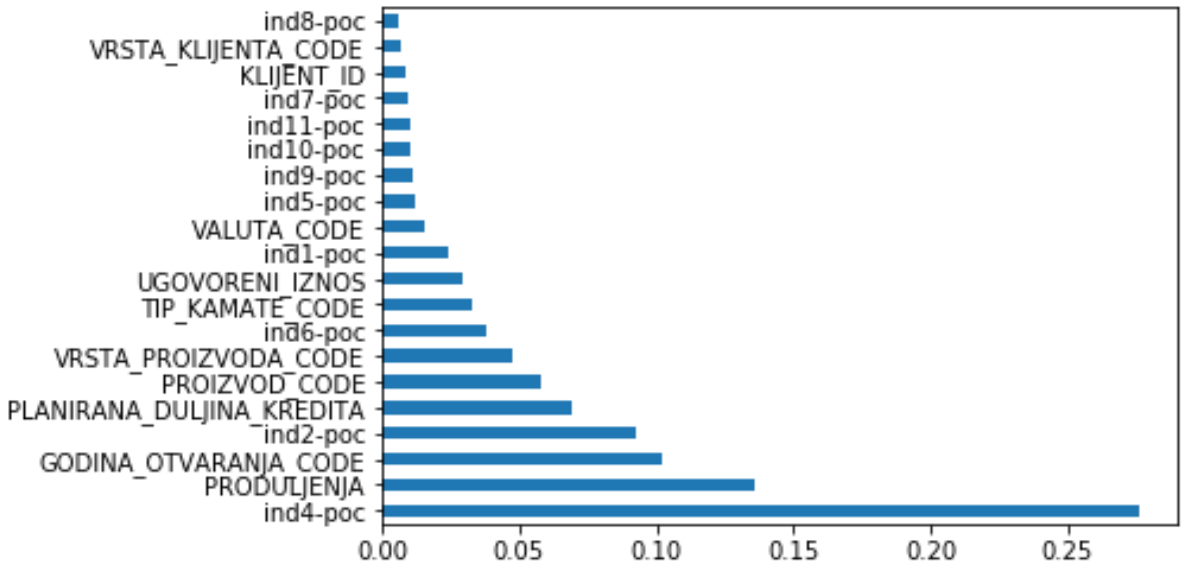
Značajka 'planirana_duljina_ugovora' (ponegdje na grafu piše 'planirana_duljina_kredita' zbog naknadnog preimenovanja) kreirana je oduzimanjem broja dana između varijabli 'planirani_datum_zatvaranja' i 'datum_otvaranja'. Značajka 'produljenja' kreirana je grupiranjem podataka po varijablama 'klijent_id' i 'oznaka_partije' i potom transformacijom po veličini za svaki datum_otvaranja' (prvi puta otvoren ugovor ima vrijednost 'produljenja' 1). Na taj način, ukupno dolazimo do 24 značajke za testiranje.

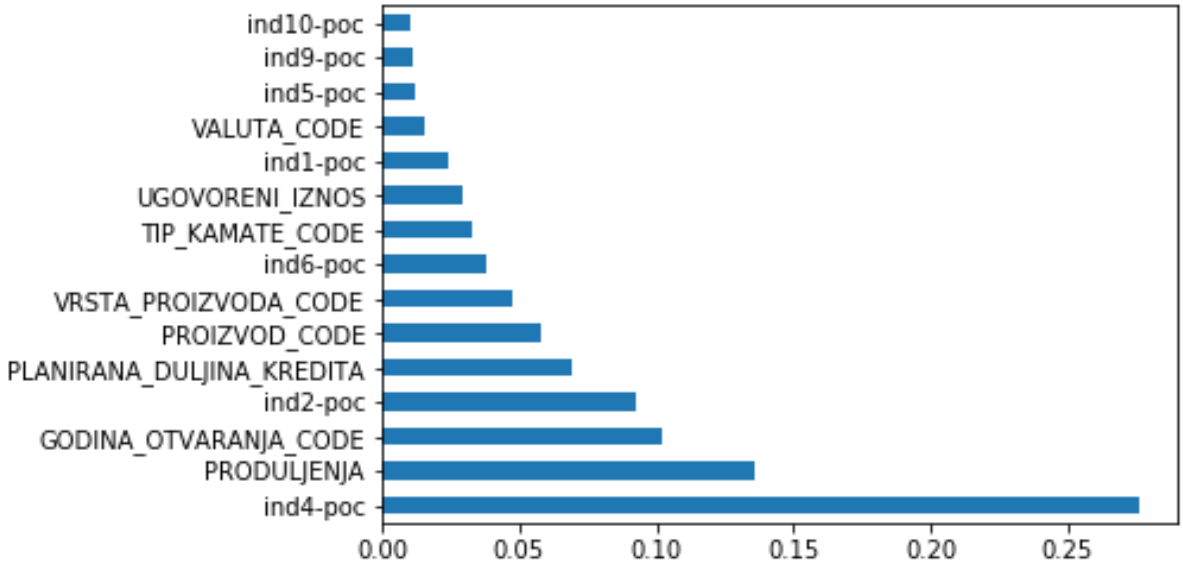
Kako bismo odredili koje ćemo značajke koristiti za treniranje modela, prvo smo iskoristili univarijatni odabir i algoritam Extremely Randomized Trees Classifier. Dobiveni su sljedeći rezultati :





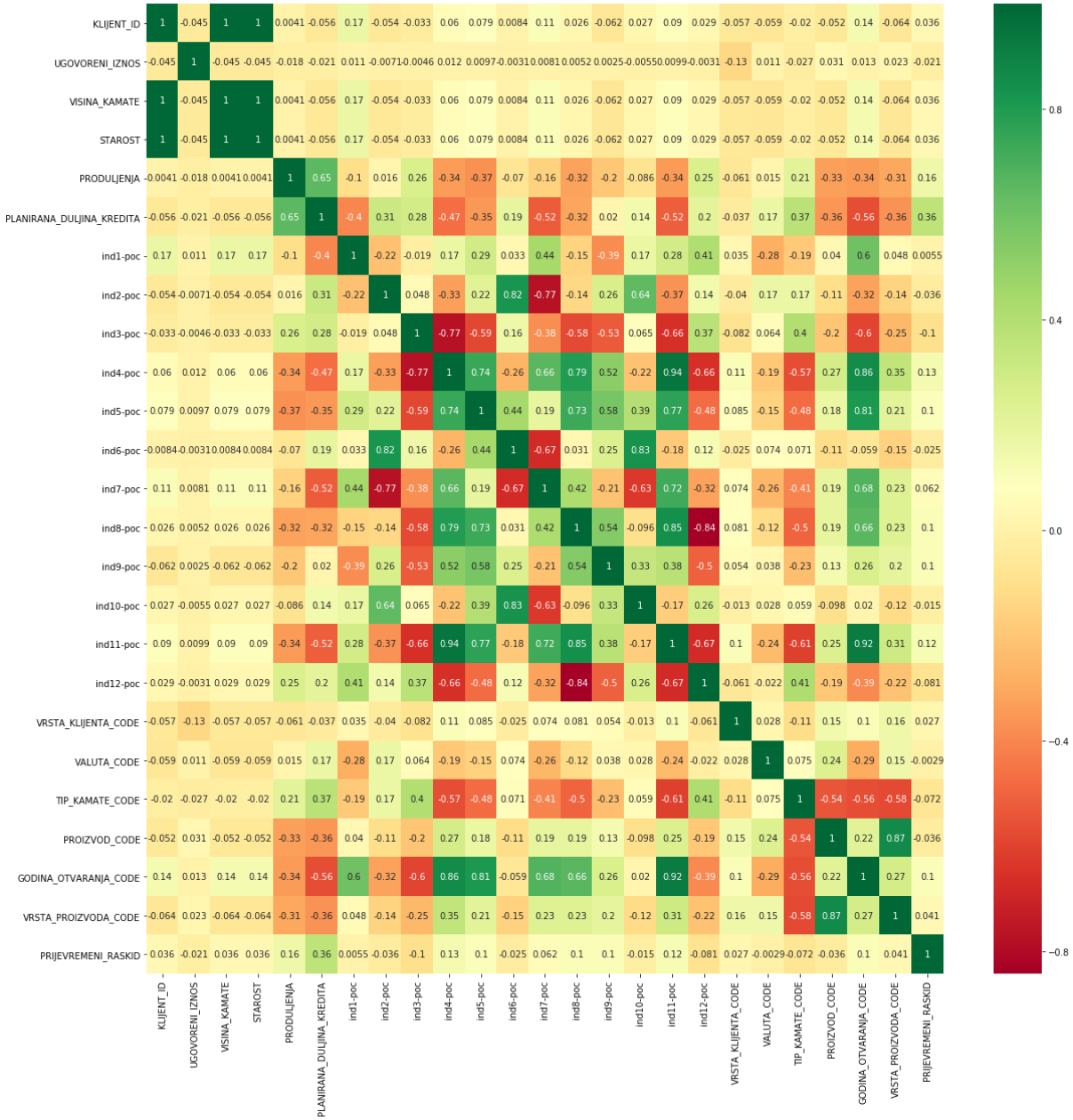
Kao bitna značajka osobito se ističe 'planirana_duljina_kredita' (ugovora). Osim toga, iskoristili smo i XGBClassifier i ponovno rangirali 15 i 20 značajki po važnosti:





U ovom slučaju, kao najznačajnije ističu se tekući račun platne bilance ('ind4-poc') te broj produljenja ugovora. Zbog neusklađenosti rezultata, odabrane su četiri različite varijante odabira značajki na kojima ćemo trenirati model (dane gornjim metodama). ('_code' označava enkodirane kategoričke varijable.)

Također, promatrali smo koreliranost značajki koristeći heatmap. Koristili smo sve 24 značajke 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.

4. Evaluirane metode

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 subsampling-a), 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 raličitim 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.

5. Analiza rješenja + future work

Tri različite metode su isprobane; slučajne šume, XGBoost te neuronske mreže, pri čemu je rađeno pretraživanje parametara za XGBoost. Od svih metoda se XGBoost pokazala kao ona koja je dala najbolje rezultate na prvom testnom skupu. Zanimljivo je to da accuracy score na pretraživanju parametara nije izgledao obećavajuće, no ROC-AUC score za istu metodu je bio znatno veći pa nam je stoga konačni rezultat bio vrlo zadovoljavajući (i f1-score i accuracy). Nedostatak metode je trajanje izvršavanja pretraživanja parametara, čak i s manjim brojem stabala.

Rezultati su bili slični za sve odabire značajki koji su dobiveni prije navedenim metodama te malo bolji od rezultata u kojem smo koristili sve 24 značajke.

Plan je bio 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 smo još razmotriti neuronske mreže za vremenske serije (LSTM) te kako bismo ih mogli iskoristiti na podacima o kvartalima.

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

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

6. Zaključak

Dani problem bio je izazovan i zanimljiv najprije zbog same strukture originalnog skupa podataka, a onda i zbog razumijevanja financijskih pojmova i praksi. Skup ima više od milijun vremenskih zapisa koje je bilo potrebno proučiti i transformirati u oblik pogodan za učenje. Također je trebalo pronaći taktiku kojom bi se popunile mnoge nelogične i nepostojeće vrijednosti. Na kraju je najbolji rezultat dao XGBoost algoritam s relativno visokom točnošću i malo višim f1 score-om. Na taj smo algoritam utrošili najviše truda i vremena, no smatramo da bi uz nastavak rada na neuronskim mrežama mogli postići još bolji rezultat. Također, informacije o kvartalima bi svakako mogle dati bolji uvid u ponašanje ugovora pa smatramo da bi SVM+ metoda mogla konkurirati metodama koje smo isprobali.