# Dokumentacija za predaju rješenja u sklopu natjecanja *Mozgalo*

***Mozgalo* 2019.**

Natjecateljski tim

Svibanj 2019.

## 1. Opis odabranog pristupa za rješavanje problema

Zadatak ovogodišnjeg natjecanja *Mozgalo* bio je predviđanje ponašanja klijenata banke *Reiffeisenbank Hrvatska*: predvidjeti hoće li se ugovor o kreditu odnosno depozitu raskinuti prije ugovorenog datuma ili ne. Taj se problem može shvatiti kao binarna klasifikacija ugovora.

Autori ovog rješenja klasifikaciji ugovora pristupili su stablima odlučivanja. Rješenje je razvijeno u programskom jeziku *Python* (inačica 3.X.X) paketom *CatBoost* (inačica 0.14.2) i sastoji se od simetričnih stabala odlučivanja, za koja implementacija *CatBoost*-a omogućuje dohvaćanje značajnosti značajki. To je svojstvo bilo poželjno zbog interpretabilnosti rješenja.

Osim standardne biblioteke *Pythona* i biblioteke paketa *CatBoost*-a, korištene su biblioteke *SciPy* paketa i paket *hyperopt*. Konačni je kôd, naravno, ipak vlastiti rukopis.

## 2. Opis *dataseta*

Osim konstrukcije samog klasifikacijskog modela, veliki je (i prvi) izazov predstavljao i sâm skup podataka — *dataset*. Prepreke su, međutim, u ovakvim problemima očekivane: otežano skupljanje točnih podataka i nedovoljno dobra prezentacija podataka analitičarima (u ovom slučaju natjecateljima).

### 2.1. *Dataset*

Sljedeće značajke dane su u treninškom *datasetu*, a podcrtane među njima dane su i u evaluacijskom i validacijskom *datasetu*:

1. *DATUM_IZVJESTAVANJA* – datum na kraju kvartala kada je izvještaj o ugovoru izrađen,
2. *KLIJENT_ID* – jedinstvena šifra klijenta čiji je ugovor bio u trenutku izvještaja,
3. *OZNAKA_PARTIJE* – jedinstvena šifra ugovora na koji se dani izvještaj odnosi,
4. *DATUM_OTVARANJA* – datum zadnjeg otvaranja (produljivanja) ugovora u trenutku izvještaja,
5. *PLANIRANI_DATUM_ZATVARANJA* – zadnji ugovoreni planirani datum zatvaranja u trenutku izvještaja,
6. *UGOVORENI_IZNOS* – zadnji ugovoreni iznos (`HRK`) ugovorenog proizvoda u trenutku izvještaja,
7. *STANJE_NA_KRAJU_PRETH_KVARTALA* – stanje ugovora (`HRK`) na kraju kvartala koji neposredno prethodi kvartalu izvještaja,
8. *STANJE_NA_KRAJU_KVARTALA* – stanje ugovora (`HRK`) na kraju kvartala izvještaja,

9. *VALUTA* – šifra valute u kojoj je proizvod ugovoren,
10. *VRSTA_KLIJENTA* – šifra vrste klijenta čiji je ugovor,
11. *PROIZVOD* – šifra ugovorenog proizvoda,
12. *VRSTA_PROIZVODA* – šifra vrste ugovorenog proizvoda (kredit ili depozit); poprima vrijednosti `'A'` i `'L'`,
13. *VISINA_KAMATE* – zadnja kamatna stopa u trenutku izvještaja,
14. *TIP_KAMATE* – šifra zadnjeg tipa kamate u ugovoru u trenutku izvještaja,
15. *STAROST* – starost klijenta (u godinama) u trenutku izvještaja — na žalost, ova je značajka irelevantna jer se na testnom skupu starost klijenta u trenutku otvaranja ili planiranog zatvaranja ugovora ne može izračunati bez datuma izvještavanja.

Trenински *dataset* sadržavao je *cca.* $5 \cdot 10^6$ ovakvih izvještaja.

Ciljna značajka koju treba predvidjeti je *PRIJEVREMENI_RASKID*: vrijednost `'Y'` za *DA* odnosno vrijednost `'N'` za *NE*. Prijevremeni raskid računa se po formuli

$$\begin{aligned}\text{PRIJEVREMENI\_RASKID}\,(\text{ugovor}) \iff \\ \iff \text{DATUM\_ZATVARANJA}\,(\text{ugovor}) + 10\,\text{dana} < \\ < \text{PLANIRANI\_DATUM\_ZATVARANJA}\,(\text{ugovor})\,, \quad (1)\end{aligned}$$

gdje uređaj $<$ na datumima znači „ranije”.

### 2.2. Vizualizacija

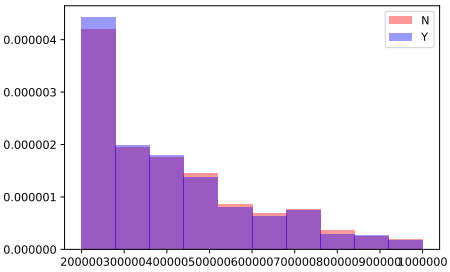
Na slici 1 prikazane su neke značajke i njihove distribucije u treninškom *datasetu* po prijevremeno raskinutim i inim ugovorima (nakon *spljoštenja dataseta* — v. dio 3). Grafovi ilustriraju kako su prijevremeno raskinuti i ini ugovori slično distribuirani, dakle, da klasifikacija nije jednostavna.

## 3. Izvlačenje značajki i njihova značajnost u korištenom modelu

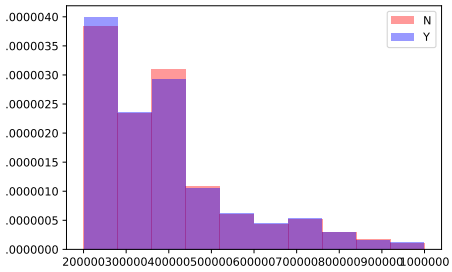
Dobiveni *dataset* (treninški, evaluacijski i validacijski) prvo se *spljošti* na način da se svi izvještaji o istom ugovoru (s istom oznakom partije) *spljošte* u jedan redak. U tom su retku zapisane vremenski prva i zadnja vrijednost značajki koje mogu varirati (na primjer, ugovoreni iznos, kamata, datumi otvaranja i planiranog zatvaranja. . . ).

Osim spljoštenja, izračunate su i neke značajke iz originalnog *dataseta*, kao, na primjer, promjena ugovorenog iznosa, duljina planiranog trajanja ugovora, kamatni račun. . . Nadalje, dodane su neke makroekonomske značajke (do datuma do kojeg postoje podatci ili predviđanja), poput prosjeka, standardne devijacije i trenda kretanja bruto domaćeg proizvoda u Republici Hrvatskoj ili cijene nafte. Konačno, dodane su neke značajke *izmišljene* vlastitom logikom — ocijene vjerojatnosti prijevremenog raskida po kombinacijama vrste klijenta, proizvoda i visine kamate izračunate statistički iz treninškog *dataseta*, obuhvaćenost perioda trajanja ugovora u krizi i slično.

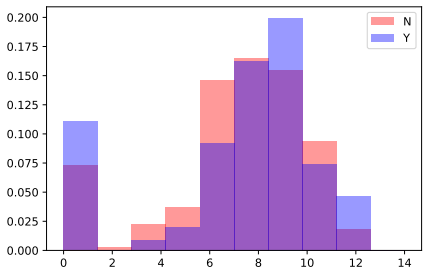
Reinterpretacijom konstruiranih modela zaključuje se da su najbitnije značajke planirani datumi zatvaranja, planirana duljina trajanja ugovora i visina kamate. Kako su konstruirana 3 modela — za ugovore vrste proizvoda `'A'` odnosno `'L'` planiranog datuma zatvaranja do 6. listopada 2016. godine i za ostale ugovore — svaki, zapravo, ima vlastite značajnosti značajki, koje se mogu vidjeti pozivom metode `get_feature_importance` na odgovarajućem modelu.



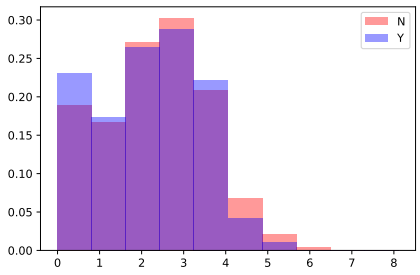

(a) `'A'`: Podjela po ugovorenom iznosu (između $2 \cdot 10^5$ i $1 \cdot 10^6$)



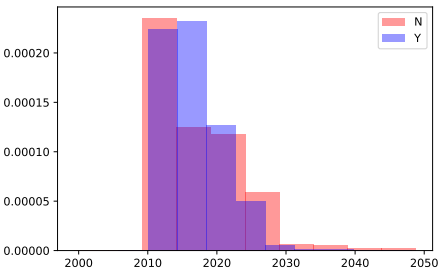

(b) `'L'`: Podjela po ugovorenom iznosu (između $2 \cdot 10^5$ i $1 \cdot 10^6$)



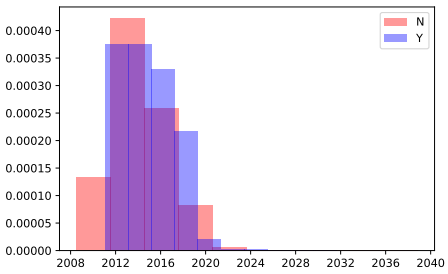

(c) `'A'`: Podjela po visini kamate (do 14)




(d) `'L'`: Podjela po visini kamate (do 8)




(e) `'A'`: Podjela po planiranom datumu zatvaranja




(f) `'L'`: Podjela po planiranom datumu zatvaranja

Slika 1: Distribucije vrijednosti nekih značajki po prijevremeno raskinutim i inim ugovorima, za svaku vrstu proizvoda posebno

## 4. Evaluirane metode

Kao što je već spomenuto u dijelu 3, konstruirana su 3 modela. Za konstrukciju svakog od modela — objekta klase `catboost.CatBoostClassifier` — paketom *hyperopt* izvršena je optimizacija hiperparametara modela, to jest, argumenata konstruktora. Optimizacija je vršena minimalizacijom razlike 1 i točnosti, preciznosti, odziva odnosno mjere $F_1$ modela (zapravo, koristila se neka težinska sredina tih vrijednosti). Odabir hiperparametara koji se optimiziraju, kao i funkcije kojima se njihove vrijednosti biraju, rezultat je odluke autora rješenja nakon čitanja dokumentacije korištenih paketa, logičnog zaključivanja i više odnosno manje uspješnih pokušaja. Nakon optimizacije hiperparametara konačno je izvršen trening na više iteracija nego koliko se koristilo u testiranju pri optimizaciji hiperparametara.

Baš zato što je ponašanje klijenata s ugovorima čiji se planirani datum zatvaranja ne će dogoditi unutar idućih nekoliko ili više godina, a konstruirani klasifikator mora predviđati trenutno stanje baze podataka banke *Reiffeisenbank Hrvatska* (umjesto stanja za 10-15 godina), ugovori čiji je planirani datum zatvaranja nakon 6. listopada 2016. godine klasificirani su kombinacijom modela za odgovarajuću vrstu proizvoda i modela za ugovore takvog kasnog planiranog datuma zatvaranja.

Datum 6. listopada 2016. godine kao granica nije odabran arbitrarno. Autori ovog rješenja zaključili su da je to datum od godinu dana nakon završetka zadnje velike krize, do kojeg makroekonomski indikatori daju kompletnu sliku vanjskih utjecaja na ponašanje klijenata, a da kasniji planirani datumi zatvaranja (pogotovo oni nakon posljednjeg datuma izvještavanja koji je 31. prosinca 2018. godine) nisu predvidivi zbog manjka adekvatne ekonomske analize ili uopće ikakvih makroekonomskih podataka.

# 5. Analiza rješenja

Autori u svom rješenju pronalaze sljedeće vrline:

1. rješenje je interpretabilno — iz modela se lako mogu *ekstrahirati* značajke na temelju kojih se donosi odluka kao i njihovi udjeli u samoj klasifikaciji,
2. rješenje generalizira problem — iako je za konstrukciju rješenja *dataset* fragmentiran na nekoliko dijelova, oni su još uvijek dovoljno veliki i općeniti da se iz njih može vidjeti logika zaključivanja, a fragmentacija modela opravdana je (v. dijelove 2.2, 3 i 4),
3. rješenje je inovativno — neke značajke izračunate su formulama koje su produkt razmišljanja autora rješenja, to jest, nisu neki općeniti ekonomski ili drugačiji indikatori kakvi se koriste u dosadašnjim modelima predikcije ponašanja klijenata, i autori smatraju da paket *CatBoost* ne uživa popularnost (korištenost) kakvu zaslužuje.

S druge strane, autori su svjesni i sljedećih mana rješenja:

1. na evaluacijskom *datasetu* mjere kvalitete prediktora bile su niže od očekivanih,
2. za treniranje modela korišten je veliki fond značajki od kojih neke možda na treninškom *datasetu* slučajno imaju korelaciju s ciljnom varijablom, što se ne može lako detektirati pozivom metode `catboost.CatBoostClassifier` (s manjim fondom značajki ta bi detekcija možda bila lakša jer bi, idealno, dominirale one značajke koje imaju opravdanu i generalizirajuću korelaciju),
3. neke su značajke međusobno inkonsistentno računate, a neke su računate nelogično ili čak pogrešno; neke pak ovise o poznatim vrijednostima makroekonomskih indikatora koji za predikciju budućnosti od nekoliko desetaka godina jednostavno ne postoje (ili postojeće predikcije donose veliku dozu nesigurnosti),
4. model pretpostavlja poznavanje *ponašanje* ugovora, to jest, je li se produljivao ili ne — ovo nužno ne mora biti mana ako je dopušteno da u predikciju ponašanja klijenta za dani ugovor bude poznato je li taj ugovor već bio ranije ugovoren od istog klijenta s drugačijim stavkama.

# Zaključak

S obzirom na dosadašnje znanje i iskustvo, kao i dostupna sredstva ((slobodno) vrijeme, računala, dostupnost podataka), autori su svojim rješenjem uglavnom zadovoljni. Ipak, preostaje dovoljno prostora za unaprjeđenje rješenja:

1. temeljitija analiza treninškog *dataseta* — zahtijeva više vremena i jače računalne kapacitete,
2. pažljivija konstrukcija dodanih značajki — nerijetko zahtijeva veću dostupnost (koja možda uopće nije moguća) vrijednosti: što dostupnih iz baze podataka banke (poput relevantne starosti klijenta, značenja vrijednosti značajki *VRSTA_KLIJENTA*, *PROIZVOD*, *TIP_KAMATE*...), što makroekonomskih pokazatelja,
3. još temeljitija analiza mogućnosti paketa *CatBoost*, kao i bolja optimizacija hiperparametara (možda i nekim drugim paketom osim paketa *hyperopt*).

Naime, različitim pokušajima, od kojih su svi vezani uz barem jednu od spomenutih točaka mogućeg daljnjeg unaprjeđenja rješenja, tijekom konstrukcije modela točnost klasifikacije rasla je, i to ponekad neočekivano velikim skokom. Autori smatraju da bi temeljitijim pristupom uvidjeli što je uspješnije pokušaje razlikovalo od neuspješnih pokušaja, čime bi se možda konstruirao zadovoljavajuće dobar model.

Međutim, autori smatraju da ekstremno kvalitetni model primjenjiv u realnom poslovanju na ovakvom problemu nije moguće konstruirati, barem ne u timovima od 3 ili 4 studenta. Za predikciju ponašanja klijenta koji u današnje vrijeme ugovara kredit ili depozit na nekoliko desetaka godina jednostavno preostaje veliki broj nepoznatih vrijednosti koje bi vrlo vjerojatno *domino-efektom* utjecale na to ponašanje. Za predviđanje tih parametara i konstrukciju modela koji je robustan na njihove greške, ako je takvo što uopće moguće, potreban je timski rad više stručnjaka iz različitih područja.