Client Behavior Prediction: A Machine Learning Challenge

Projektni prijedlog - Mozgalo

Elena Murljačić, Dorotea Rajšel, Darija Strmečki, Petra Vlaić

Travanj 2019.

1 Uvodni opis problema

Pri sklapanju ugovora između banke i klijenta, banka razmatra razne faktore koji bi mogli utjecati na ponašanje klijenta. Ponašanje klijenta 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.

Zadatak je zadan na studentskom natjecanju Mozgalo. Na raspolaganju su 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 slično, što numeričkih, što kategoričkih. Problem je klasifikacijski; ciljna značajka je kategorička varijabla koja govori je li ugovor (kredit ili štednja) prijevremeno raskinut ili ne. Navedene podatke RBA već koristi u simulacijama za predviđanje zatvaranja kredita tako da će učenje iz tih podataka biti izvedivo.

2 Cilj i hipoteze istraživanja problema

U interesu je banke pokušati predvidjeti klijente koji će potencijalno promijeniti ugovoreni odnos s bankom u smislu prijevremene otplate kredita ili produženja oročenog depozita. Temeljem dobivene analize, potrebno je pripremiti nove presonalizirane ponude za klijente. Cilj je prilagoditi ponudu predviđenim novonastalim uvjetima te na taj način umanjiti negativne efekte na financijski rezultat.

3 Pregled dosadašnjih istraživanja

Budući da je zadatak zadan na studentskom natjecanju Mozgalo, ne postoje prethodna istraživanja i rezultati koji koriste isti dataset, a na koje bismo se mogli referirati. Međutim, razmotrili smo neke slične probleme vezane za procjenjivanje kreditnog rizika, "customer churn prediction" problema i općenito problema s podacima u obliku vremenskih nizova.

3.1 Machine Learning in credit risk modeling - tvrtka James

U radu [6], američka tvrtka James koristi tipični dataset za kreditni rizik od približno 150,000 zapisa sa 12 značajki i predviđa je li kredit "default" ili "safe". U podatcima se eliminiraju kolinearne značajke te se po potrebi koristi data binning. Nakon pretprocesiranja podataka, koristi se cross validation za treniranje modela na jednom dijelu dataseta (80%), a testira se na preostalom dijelu (20%). Tvrtka je odabrala 3 modela - Logističku regresiju, Random Forest i Gradient Boosting. Dobiva se da je ROC-AUC score za Logističku regresiju 0.75, za Random Forest je 0.83 te za Gradient Boosting 0.84.

3.2 Machine learning techniques for customer churn prediction in banking environments, Valentino Avon

Diplomski rad [5] promatra dataset u kojem za određenog klijenta postoji više zapisa (konkretno za svaki mjesec). Dataset se transformira smanjivanjem dimenzije i uvođenjem tzv. "multi-month" atributa da bi se dobio jedan zapis po klijentu te tako moglo lakše baratati podacima. U ovom radu uspoređivala se točnost i vrijeme potrebno za treniranje. Dobiva se da najbolji omjer točnosti i vremena ima algoritam C5.0 koji je korišten u kombinaciji s undersampling tehnikom.

3.3 Time Series Forecasting with RNNs, Marek Galovič

U članku [12] dan je pregled RNN modela za predviđanje s podacima koji su u obliku vremenskih nizova. Zaključuje se da su neuronske mreže moćan alat za modeliranje vremenskih nizova i mogu se primijeniti na cijelom spektru problema, od predviđanja prodaje do predviđanja potrošnje električne energije. Kao primjer u članku je dano predviđanje cijene dionica.

4 Materijali, metodologija i plan istraživanja

4.1 Pristup rješavanju problema

Problem ćemo pokušati riješiti koristeći metode nadziranog učenja za klasifikaciju. Trenutačno na raspolaganju imamo samo skup podataka za treniranje. Kada nam Mozgalo ustupi skup podataka za testiranje, na njemu ćemo također provesti metode procjene preciznosti i validacije modela.

4.2 Prikupljanje podataka

Podatke o transakcijskim računima pravnih i fizičkih osoba ustupila je RBA. Za bolju klasifikaciju, planiramo iskoristiti podatke o makroekonomskim pokazateljima sa stranica Državnog zavoda za statistiku [2] i Hrvatske narodne banke [1] (kao što su prosječna plaća, BDP, inflacija, državni dug i stopa zaposlenosti).

Naš 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.

Svaki zapis nam predstavlja izvještaj o ugovoru (kreditu ili štednji) određenog klijenta u jednom kvartalu, dakle imamo više vremenskih nizova za jedan kredit. U eksplorativnoj analizi, grupiranjem smo dobili da postoji ukupno 950239 različitih ugovora. Naši korisnici mogu biti pravne ili fizičke osobe, no to nije vidljivo u originalnom skupu podataka te bi to trebalo na neki način odrediti zbog lakše analize.

4.3 Metode i algoritmi

Podatke ćemo analizirati koristeći programski jezik Python unutar Jupyter Notebook okruženja. Odabrani su zbog jednostavnosti vizualizacije podataka i implementacije raznih korisnih biblioteka i alata, kao što su Pandas, Numpy, Graphviz, Sklearn i slično.

Preoblikovat ćemo skup podataka jer po jednom primjeru imamo više vremenskih nizova, tj. u skupu se nalaze izvještaji po kvartalima za svakog klijenta i svaki njegov kredit/štednju. Kako bismo mogli učiti modele na tom skupu, pomoću tih vremenskih nizova želimo definirati varijable kojima ćemo opisati trend plaćanja kroz cijelo vremensko razdoblje.

Skup podataka je nebalansiran (otprilike 16% ugovora je prijevremeno raskinuto) pa je jedan od problema koji očekujemo svakako overfitting. Na radionicama Mozgala savjetovano nam je umanjiti taj problem koristeći unakrsnu validaciju i tehnike sampliranja. Za treniranje modela istražit ćemo slučajne šume, ansambl metode i neuronske mreže (ovisno o vremenu) zbog toga što su pogodne za takve skupove.

Da bi se odabir značajki i treniranje modela moglo provesti, potrebno je prije napraviti enkodiranje kategoričkih značajki. Nakon toga, za odabir značajki koristit ćemo metode iz sklearn.feature_selection. Prvo ćemo provesti univarijatni odabir, da dobijemo bolji uvid u važnost pojedine značajke. Očekujemo da neke značajke ovim odabirom neće biti dosta informativne, nego će u paru s nekim drugima imati veći značaj pa ih stoga nećemo odmah odbaciti. Promotrit ćemo i međusobnu koreliranost varijabli. Nadalje, koristit ćemo metodu slučajnih šuma pa će se time izravno računati važnost značajki. Također, isprobat ćemo i rekurzivnu eliminaciju značajki s unakrsnom validacijom (pomoću slučajnih šuma).

4.4 Ocjena uspješnosti rezultata

Prilikom izgradnje modela, za ocjenu uspješnosti koristit ćemo dvije mjere: preciznost (accuracy) i F1 score. Konačno rješenje bit će analizirano i ocijenjeno od strane Mozgala: najbitnija mjera bit će preciznost, a F1 score će presuditi u slučaju da više timova postigne jednaku točnost.

Adacta će osigurati web servis putem kojeg će natjecatelji moći uploadati rezultate svojih modela i jednom dnevno bit će moguće vidjeti trenutačni raspored timova uzimajući u obzir preciznost rezultata treniranog modela. Barem jednom dnevno ćemo na taj način evaluirati naš model.

5 Očekivani rezultati predloženog projekta

Podatci na kojima vršimo analizu su jedinstveni te ne postoje prijašnja istraživanja. Iz tog razloga, teško je dati realnu procjenu uspješnosti za naš skup. Promatrajući istraživanja sličnih problema, najuspješniji je bio ROC-AUC score od 0.84 iz rada [6] pa se nadamo približno jednakom uspjehu, ako ne i boljem.

Literatura

[1] URL: https://www.hnb.hr/statistika/glavni-makroekonomski-indikatori.

[2] URL: https://www.dzs.hr/.

[3] URL: https://www.coursera.org/learn/machine-learning.

[4] URL: http://stakana.com/.

[5] Valentino Avon. „Machine learning techniques for customer churn prediction in banking environments". Mag. rad. Universit'a degli Studi di Padova, 2016.

[6] Leonardo Baldassini i Jose Antonio Rodriguez Serrano. „client2vec: Towards Systematic Baselines for Banking Applications". (2018).

[7] Olivier Blanchard. *Macroeconomics*. 3rd. 2015.

[8] Nate Derby. „Maximizing Cross-Sell Opportunities with Predictive Analytics for Financial Institutions". *Paper 941-2017* (2017).

[9] Nate Derby i Mark Keintz. „Reducing Customer Attrition with Predictive Analytics for Financial Institutions". *MWSUG 2016 - Paper BI04* (2016).

[10] D.W. Findlay i Olivier Blanchard. *Macroeconomics-Study Guide*.

[11] Jerome H. Friedman, Robert Tibshirani i Trevor Hastie. *Elements of Statistical Learning*. 2008.

[12] Marek Galovič. „Time Series Forecasting with RNN's". (2018).

[13] J. D. Hamilton. *Time Series Analysis*.

[14] Tomislava Pavić Kramarić. *Osnove financija*.

[15] R.L. Thomas. *Modern econometrics*.