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 serija. Za svaku fizičku/pravnu osobu dostupno je 17 parametara kao što su ugovoreni iznos, datum otvaranja, visina kamate i slično. 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 i "customer churn prediction" problema.

3.1 Machine Learning in credit risk modeling, tvrtka James

U ovom radu, američka tvrtka James koristi tipični dataset za kreditni rizik od približno 150,000 zapisa sa 12 feature-a i predviđa je li kredit "default" ili "safe". U podatcima se eliminiraju kolinearni featuri 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. Dobili su 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

Ovaj diplomski rad 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. Dobiveno je da najbolji omjer točnosti i vremena ima algoritam C5.0 koji je korišten u kombinaciji s undersampling tehnikom.

4 Materijali, metodologija i plan istraživanja

4.1 Pristup rješavanju problema

Problem ćemo pokušati riješiti koristeći metode nadziranog učenja. 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 predikciju, planiramo iskoristiti podatke o makroekonomskim pokazateljima sa stranica Državnog zavoda za statistiku i Hrvatske narodne banke (kao što su prosječna plaća, BDP, inflacija, državni dug i stopa zaposlenosti).

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 pa je jedan od problema koji očekujemo svakako overfitting. Na radionicama Mozgala savjetovano nam je umanjiti taj problem koristeći cross validation i data augmentation. Za treniranje modela istražit ćemo slučajne šume, ansambl metode i neuronske mreže (ovisno o vremenu).

4.4 Ocjena uspješnosti rezultata

Prilikom izgradnje modela, za ocjenu uspješnosti koristit ćemo F1 score. Konačno rješenje bit će analizirano i ocijenjeno od strane Mozgala.

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. Pokušat ćemo dobiti model koji daje točnost od barem 90% na testnom skupu podataka.

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

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

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

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

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

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

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

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