Predviđanje uspjeha bankarskog marketinga

Projektni prijedlog za kolegij Strojno učenje

Davorin Gradečak, Petra Jambriško, Jakov Krunić

17. travnja 2020.

Sadržaj

1 Uvodni opis problema

Problem kojim se bavimo je predviđanje uspjeha bankarskog marketinga na temelju popisnih podataka. Marketinške kampanje predstavljaju tipičnu strategiju unapređenja određenog poslovanja. Jedna od čestih marketinških metoda je telemarketing—interakcija s korisnicima na daljinu putem telefona.
Podaci koje koristimo su nastali u marketinškim kampanjama koje su se bazirale na telefonskim pozivima. Banke su direktno kontaktirale svoje korisnike s ciljem da korisnik ugovori oročeni depozit. Korisnici su mogli biti kontaktirani i više puta.
Podaci su podijeljeni u dva skupa podataka. To su *train* podaci i *test* podaci. Skup podataka *train* se sastoji od 41188 instanci, a skup podataka *test* se sastoji od 4119 instanci, tj. 10% instanci iz *train* skupa koje su odabrane nasumično. Svaka instanca predstavlja određenog korisnika banke. Svaki korisnik je opisan s 20 atributa i atributom y koji predviđamo (target). Atributi su vezani uz samog korisnika (godine, spol, vrsta posla ...), uz zadnji kontakt banke s korisnikom tijekom kampanje (dan u tjednu, trajanje ...), uz prethodne kontakte s korisnikom (broj kontakata prije i tijekom trenutne kampanje ...) i uz neke ekonomske podatke tijekom trenutne kampanje (indeks potrošačkih cijena ...). Atribut y nam govori je li korisnik ugovorio oročeni depozit ili ne. Dakle y je zapravo kategorijska varijabla s dvije kategorije "yes" i "no".

2 Cilj i hipoteze istraživanja

Jedan od ciljeva našeg istraživanja je odrediti utječu li pojedini atributi iz skupa podataka na varijablu y koju predviđamo (target). Drugim riječima, zanima nas koliki je utjecaj određenih karakteristika korisnika na njegovu odluku o tome hoće li ugovoriti oročeni depozit ili ne.
Osim toga, nastojimo napraviti što uspješniji binarni klasifikator, tj. model koji što uspješnije predviđa hoće li korisnik ugovoriti oročeni depozit. U tu svrhu ćemo implementirati i testirati neke algoritme strojnog učenja te ih na kraju evaluirati na temelju tehnika opisanih u odjeljku 4.4 *Ocjena uspješnosti rezultata projekta* i usporediti. Hipoteza našeg istraživanja jest: moguće je predvidjeti odluku korisnika o tome hoće li ugovoriti oročeni depozit na temelju raznih atributa. Drugim riječima, tvrdimo da je moguće primjenom adekvatnih algoritama strojnog učenja dobiti model koji uspješno predviđa odluku korisnika.

3 Pregled dosadašnjih istraživanja

U radu [3] korištene su četiri klasifikacijske metode. To su logistička regresija, stablo odlučivanja, neuronske mreže i metoda potpornih vektora. Ove metode su uspoređene pomoću dvije klasifikacijske metrike: AUC (Area Under Curve) i ALIFT (Area of the LIFT cumulative curve). Za obe metrike, najbolji rezultati su ostvareni metodom neuronskih mreža (AUC = 0.8 i ALIFT = 0.67). Korišten je paket *rminers* programskog jezika R.

Rad [1] se bazira na usporedbi više različitih algoritama: algoritam slučajnih šuma, metoda potpornih vektora, neuronske mreže, naivni Bayesov algoritam i algoritam k-najbližih susjeda. Rezultati ovog rada su pokazali da se najbolja predviđanja postižu metodom potpornih vektora i naivnim Bayesovim algoritmom. Zbog jednostavnosti i održivosti, preferira se naivni Bayesov algoritam.
U ostalim radovima koje smo pronašli uglavnom se javljaju već nabrojani algoritmi. Također, mnoga su dosadašnja istraživanja koristila samo dio skupa podataka, dok je u našem planu iskoristiti što više dostupnih podataka.

4 Materijali, metodologija i plan istraživanja

4.1 Način rješavanja problema

Budući da je vrijednost koju naši algoritmi trebaju predvidjeti binarna varijabla, problem koji obrađujemo je klasifikacijski. Problem ćemo rješavati primjenom metoda i algoritama nadziranog učenja. Ispitat ćemo efikasnost više različitih algoritama te ih međusobno usporediti. Na kraju ćemo dati zaključak o tome koji je algoritam bio najtočniji. Navest ćemo i smjernice za eventualna nova istraživanja koja bi se nadovezala na naš rad.

4.2 Prikupljanje podataka

Podatke preuzimamo s web stranice UCI Machine Learning Repository (novija verzija podataka u [2]). Podaci se odnose na izravne marketinške kampanje portugalske bankarske institucije. Skup podataka koji ćemo koristiti je gotovo identičan onome koji je korišten u [3].

4.3 Metode/algoritmi/tehnike/alati

U svrhu rješavanja problema planiramo koristiti programski jezik Python te raznovrsne njegove biblioteke kao što su Numpy, Pandas, Scipy, Matplotlib, Scikit-learn itd. Kao razvojno okruženje koristit ćemo Jupyter bilježnicu. Za rješavanje problema planiramo koristiti algoritme nadziranog učenja kao što su logistička regresija, neuronske mreže, naivni Bayes, k-najbližih susjeda, slučajne šume te metoda potpornih vektora. Na nekoliko različitih načina preuredit ćemo naše podatke te uspoređivati uspješnost algoritama s obzirom na različite algoritme i uređivanje podataka. Prilikom razvoja algoritama koristit ćemo PCA (Analizu glavnih komponenata) te vidjeti utječe li taj postupak na poboljšanje rezultata. Također, pri analizi rezultata promatrat ćemo razliku uspješnosti algoritama s obzirom na način rješavanja problema nedostajućih vrijednosti.

4.4 Ocjena uspješnosti rezultata projekta

Model ćemo primarno evaluirati pomoću vrijednosti AUC (Area Under Curve), tj. površine ispod ROC (Receiver Operating Characteristic) krivulje. To je krivulja koja

prikazuje odnos TPR (true positive rate) u odnosu na FPR (false positive rate). Pri tome je TPR broj korektnih klasifikacija u (pozitivnoj) klasi u odnosu na ukupan broj pozitivnih primjera, a FPR je definiran kao broj krivih klasifikacija u (pozitivnoj) klasi u odnosu na ukupan broj negativnih primjera. Cilj je približiti se vrijednosti savršenog klasifikatora (AUC = 1). Osim toga, odredit ćemo matricu konfuzije, vrijednost preciznosti P i F_1—mjeru koja povezuje preciznost i osjetljivost.

5 Očekivani rezultati predloženog projekta

Kao konačni rezultat projekta očekujemo model koji što uspješnije rješava dani klasifikacijski problem. Cilj nam je napraviti model sa što većom vrijednošću AUC. Očekujemo dobiti preciznost kao i drugi ili bolje od toga. Naše rezultate ćemo usporediti s već poznatim rezultatima.

Literatura

[1] A. Abu-Srhan i dr. *Visualization and Analysis in Bank Direct Marketing Prediction*. 2019. URL: https://thesai.org/Downloads/Volume10No7/Paper_85-Visualization_and_Analysis_in_Bank_Direct_Marketing.pdf.

[2] *Bank Marketing Data Set*. http://archive.ics.uci.edu/ml/datasets/Bank+Marketing. [Zadnje pristupljeno 17.4.2020].

[3] S. Moro, P. Cortez i P. Rita. *A data-driven approach to predict the success of bank telemarketing*. 2014. URL: http://media.salford-systems.com/video/tutorial/2015/targeted_marketing.pdf.