Prediktivno određivanje uspješnosti telemarketinga banaka

Projektni prijedlog kolegija Strojno učenje, ak.god. 2017./2018.

Prirodoslovno-matematički fakultet Sveučilišta u Zagrebu

Profesor: dr. sc. Tomislav Šmuc

Asistenti: Tomislav Lipić, Matej Mihelčić, Matija Piškorec

Studenti: Jelena Krnjak i Tomislav Smetko

travanj, 2018.

1. Opis problema

Cilj poslovnih banaka je ostvarenje profita. Banke, u pravilu, ostvaruju profit kroz zaradu od kamata koje klijenti plaćaju na kredite. Da bi mogle plasirati kredite, banke moraju imati kapital, koji uglavnom ostvaruju primajući oročene depozite klijenata. Od velike financijske krize 2008. godine, od banaka se traži da imaju velike rezerve novca kojima pokrivaju izdane kredite. Zato banke žele klijente koji će dugoročno oročiti depozite u njihovoj banci. Kako bi privukle što više klijenata da dugoročno oroče depozite, koriste se marketinške kampanje u različitim oblicima. U ovom radu promatrat ćemo oblik marketinške kampanje telemarketinga – metode u kojoj banka organizira pozivni centar iz kojeg upućuje pozive klijentima te im pokušava prodati uslugu-u ovom konkretnom slučaju nagovoriti ih na dugoročno držanje depozita u banci.

2.1. Cilj i hipoteze istraživanja

Cilj istraživanja je odrediti klasifikacijski model koji će za zadanu instancu (klijenta banke) na osnovu toj instanci pridruženih značajki odrediti hoće li odabrani klijent dugoročno oročiti depozit u banci, prije uspostve samog poziva od strane pozivng centra. Dakle, želja je da banka uspije ciljano kontaktirati one klijente za koje je veća vjerojatnost da će dugoročno oročiti depozite. Glavne prednosti ovakvog ciljanog komuniciranja je smanjenje troška (za banku) i smanjenje invazivnosti na klijente.

3. Pregled dosadašnjih istraživanja

Prvi koji su se bavili ovim problemom bili su Moro i suradnici ([1]). Dataset koji se koristi u njihovom radu koristit će se i u ovom projektu, no sa smanjenim brojem značajki. U originalnom radu autori su krenuli sa 150 značajki, no korištenjem metoda redukcije dimenzionalnosti sveli su originalni dataset na 21 najvažnijih značajki (koje će se koristiti u ovom radu). Da bi riješili problem koristili su metode logističke regresije (LR), metode potpornjih vektora (SVM) i neuronskih mreža (NN). Istraživanje je pokazalo da najbolje rezultate daju NN. Ovom metodom utvrđeno je da 79% pozitivno realiziranih kontakata sa klijentima može biti izrealizirano kontaktiranjem samo polovine klijenata, što je znatno poboljšanje s obzirom na standardnu metodu kontaktiranja svih klijenata.
Kasnije su ovi rezultati ponovljeni u još nekim radovima ([2], [3]) te je korištena metoda random forest (RF) kojom su dobiveni slični rezultati kao i u [1].

4.1. Materijali

Podatci koji se koriste su dio dataseta *Bank Marketing Data Set* preuzetog sa repozitorija Sveučilišta Irvine u Kaliforniji ([4]). Dataset se sastoji od 45212 instanci koje predstavljaju klijente banke kojima je pridruženo 21 značajki. Podatci su prikupljeni u razdoblju između 2008. i 2013. godine (dakle podatcima je pokriveno razdoblje velike financijske krize) od jedne portugalske poslovne banke. Svaku od značajki možemo svrstati u jednu od pet kategorija: demografski podatci o klijentu (DP), klijentove financijske karakteristike (FK), podatci vezani uz trenutnu marketinšku kampanju banke (TMK), podatci vezani uz prethodnu marketinšku kampanju (PMK) i socioekonomska situacija (SES). Detaljan popis značajki s opisom, tipom te kategorijom kojoj pripada i rasponom vrijednosti koje poprima dostupni su u tablici 4.1.1.
Napomena: podatci o socioekonomskoj situaciji preuzeti su sa službene web stranice portugalskog Državnog zavoda za statistiku.

#	značajka	opis	tip	kategorija	vrijednosti
1	age	klijentova dob	numerički	DP	[17, 98]
2	job	vrsta klijentovog posla	kategorički	DP	{admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown}
3	marital	bračni status	kategorički	DP	{divorced, married, single, unknown} (divorced means divorced or widowed)
4	education	stupanj obrazovanja klijenta	kategorički	DP	{basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown}
5	default	klijentova sklonost dovođenja kreditnih kartica do limita preko kojega klijent nije više u stanju vratiti dug (defaulta)	kategorički	FK	{no, yes, unknown}
6	housing	ima li klijent stambeni kredit	kategorički	FK	{no, yes, unknown}
7	loan	ima li klijent nenamijenski kredit	kategorički	FK	{no, yes, unknown}
8	contact	način posljednjeg uspostavljanja kontakta s bankom	kategorički	TMK	{cellular, telephone}
9	month	naziv mjeseca u kojem je ostvaren posljednji kontakt	kategorički	TMK	{jan, feb, mar, ..., nov, dec}
10	day_of_week	naziv dana u tjednu u kojem je ostvaren posljednji kontakt	kategorički	TMK	{mon, tue, wed, thu, fri}
11	duration	trajanje posljednjeg poziva (u sekundama)	numerički	TMK	[0,4918]
12	campaign	broj kontakata s klijentom ostvarenih u trenutnoj kampanji (uključujući trenutni poziv)	numerički	TMK	[1, 56]
13	pdays	broj dana proteklih od posljenjeg kontakta s klijentom u trenutnoj kampanji	numerički	PMK	[0, 27], {999} (999 znači da klijent nije prethodno kontaktiran)
14	previous	broj kontakata s klijentom do početka trenutne kampanje	numerički	PMK	[0,7]
15	poutcome	ishod prethodne kampanje	kategorički	PMK	{failure, nonexistent, success}
16	emp.var.rate	kvartalni indicator stope zaposlenosti	numerički	SES	[-3.4,1.4]
17	cons.price.idx	mjesečni indikator cijena	numerički	SES	[92.201,94.767]
18	cons.pconf.idx	mjesečni indikator korisničkog optimizma	numerički	SES	[-50.8,-26.9]
19	euribor3m	tromjesečna Euribor kamatna stopa	numerički	SES	[0.634,5.045]
20	nr.employed	kvartalni indikator broja zaposlenih	numerički	SES	[4963.6,5228.1]
21	subscription	da li klijent ima dugoročno oročeni deposit u banci	kategorički	FK	{yes, no}

Tablica 4.1.1. Popis značajki dataseta

4.2. Metodologija i plan istraživanja

Problem ćemo pokušati riješit korištenjem algoritama i metoda strojnog učenja. Usporediti ćemo nekoliko metoda, odnosno algoritama, te tako pokazati koji od tih modela bi bio najbolji izbor kao riješenje.

Modele koje ćemo uspoređivati su Naivni Bayesov kvantifikator (NB), Stabla odlučivanja (DT), slučajne šume (RF) te stroj s potpornim vektorima (SVM).

Kao alat koristiti ćemo se implementacijam modela iz Pythonove biblioteke scikit-learn [8].
Za NB ćemo koristiti verziju Gaussovog naivnog Bayesa implementiranu u funkciji *sklearn.naive_bayes.GaussianNB* za klasifikaciju u prostoru numeričkih značajki [9]. Slučajne šume su implementirana u *sklearn.ensemble.RandomForestClassifier* [10], dok je SVN implementiran u klasi *sklearn.svm.SVC* [11], a DT u klasi *sklearn.tree.DecisionTreeClassifier* [12].

Pošto se ovdje radi o nebalansiranim podacima za metriku smo odabrali Receiver Operating Characteristic (ROC).

Kao konačni rezultat planiramo napraviti usporedbu između modela strojnog učenja koji ćemo korisiti za riješenje problema, te usporedba tih rezultata s orginalnim radom [1].

Literatura

[1] Moro, S., Cortez, P., Rita, P.: A Data-Driven Approach to Predict the Success of Bank Telemarketing,2014.

[2] Ejaz, S.: Predicting Demographic and Financial Attributes in a Bank Marketing Dataset,2016., diplomski rad

[3] Kim, K., Lee, C., Jo, S.,Cho, S.: Predicting the Success of Bank Telemarketing using Deep Convolutional Neural Network,2015.

[4] Dataset: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[5] https://github.com/lingamjetta/Success-of-Bank-Telemarketing-System

[6] https://danielabban.github.io/2017/04/predicting-the-success-of-bank-telemarketing/

[7] https://danielabban.github.io/2017/04/predicting-the-success-of-bank-telemarketing/

[8] http://scikit-learn.org/stable/index.html

[9] http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

[10]http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[11] http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC

[12] http://scikit-learn.org/stable/modules/tree.html