Prediktivno određivanje uspješnosti telemarketinga banaka

Jelena Krnjak

Tomislav Smetko

Projekt kolegija Strojno učenje, lipanj 2018.

Prirodoslovno-matematički fakultet Sveučilišta u Zagrebu

Sažetak

U ovom radu izrađeni su klasifikator za određivanje uspješnosti telemarketinga banke čiji je cilj prikupljanje oročenih depozita klijenata. Dataset korišten u ovom radu je Bank Marketing Data Set preuzet sa repozitorija Sveučilišta Irvine u Kaliforniji, na kojem je do sada proveden veći broj razmatranja, što daje priliku za usporedbu razultata. Za razliku od originalnog istraživanja, u kojem je dataset imao 150 značajki, dataset korišten u ovom radu imao je 20 značajki (plus output varijabla). Algoritmi strojnog učenja korišteni u ovom radu su oni algoritmi koji su u literaturi dali najbolje rezultate: naivni Bayes, stable odlučivanja, slučajne šume, stroj s potpornim vektorima (SVM).

Najbolji rezultati dobiveni su algoritmom SVM, a najmanje dobri rezultati metodom slučajnih šuma.

Uvod

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 promatran je 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.

Cilj istraživanja je određivanje klasifikacijskog modela 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 odrediti (kontaktirati) one klijente za koje je velika vjerojatnost da će dugoročno oročiti depozite. Glavne prednosti ovakvog ciljanog komuniciranja je smanjenje troška promocije (za banku) i smanjenje invazivnosti na klijente.

Skup podataka i prikupljanje

Podatci koji se koriste su dio dataseta *Bank Marketing Data Set* preuzetog sa repozitorija Sveučilišta Irvine u Kaliforniji ([9]). Dataset

#	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 1. Popis značajki dataseta

se sastoji od 41188 instanci koje predstavljaju klijente banke kojima je pridruženo 20 značajki (ne uključujući output varijablu). 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 1.

Napomena: podatci o socioekonomskoj situaciji preuzeti su sa službene web stranice portugalskog Državnog zavoda za statistiku.

Za svaku kategoričku varijablu i za svaku njezinu moguću vrijednost u datasetu dodan je novi stupac. Novi stupac za neku kategoričku varijablu i njezinu vrijednost je 1 ako orginalna varijabla ima tu vrijednost, a 0 inače. Numeričke varijable normalizirane su na intervalu [0,1].

Dataset je na slučajan način podijeljen u trening set (skup primjera na kojima se pojedini model trenira) i test set (skup primjera na kojima se ocjenjuje uspješnost modela).

Algoritmi strojnog ućenja

Modeli koji će biti uspoređivani u ovom radu i su Naivni Bayes, stabla odlučivanja, slučajne šume te stroj s potpornim vektorima (SVM).

Naivni Bayes

Prvi korišteni algoritam je Naivni Bayes.

Naivni Bayes je model učenja koji koristi Bayesov teorem za procjenu vjerojatnosti pojedine klase. Osnovna pretpostavka algoritma je da su značajke nezavisne, što je rijetkost u praksi, no ova metoda daje dobre rezultate.

Ovim algoritmom dobivena je točnost T (eng. accuracy), broj koji se računa kao

$$T = \frac{TP+TN}{TP+FP+TN+FN},$$

gdje su

TP - broj stvarno pozitivnih primjera, točno predviđenih od strane modela

TN-broj stvarno negativnih primjera, koji su točno predviđeni od strane modela kao negativni

FP- broj stvarno negativnih primjera, koji su netočno predviđeni od strane modela kao pozitivni

FN- broj stvarno pozitivnih primjera, koji su netočno predviđeni od strane modela kao negativni

Dobiveni su sljedeći podatci:

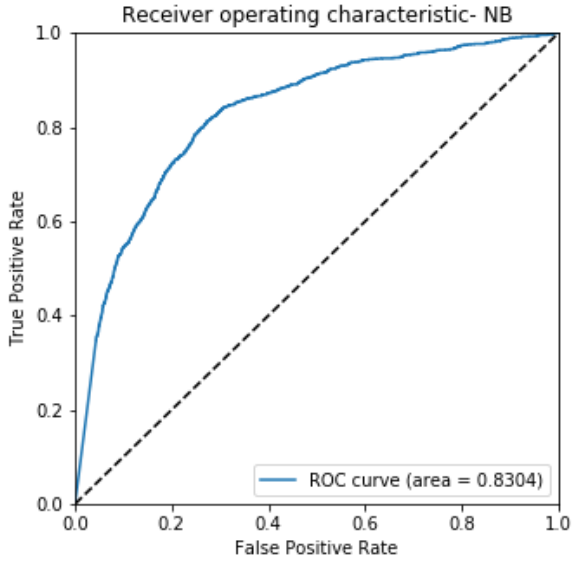
T= 0. 8045317442801442,

a točnost na test setu je dobivena

Ttest= 0.7987375576596262.

Zbog nebalansiranosti podataka, kao mjera uspješnosti odabrana je tzv. ROC krivulja (Receiver Operating Characteristic), koja prikazuje odnos TP i FP. Pri tome,promatra se površina ispod krivulje (eng. area under curve, AUC) te vrijedi da je klasifikator to bolji što je površina ispod ROC krivulje veća. Dakle, za savršeni klasifikator vrijedi AUC=1, dok onaj klasifikator za kojeg vrijedi da je AUC=0.5 slučajno klasificira primjere.

ROC krivulja za naivni Bayes ovog dataseta prikazana je na slici 1.



Slika 1.: ROC krivulja, Naivni Bayes

Dobiva se da je, za naivni Bayes algoritam promatranog dataseta dobivena sljedeća vrijednost AUC:

AUC=0.8304

AUCtest=0.8499 (za test set)

Prema ovom modelu, banka će telemarketinškom kampanjom kontaktirati 83% željenih klijenata 8onih koji će dugoročno oročiti depozite).

Stabla odlučivanja

Stablo odluke (eng. decision tree) je algoritam koji modelira proces donošenja odlukate njihovih posljedica. Učenje stablom odluke koristi stabla odluke kao prediktivni model s ciljem da se iz dostupnih informacija o nekom predmetu (prikazanih pomoću grana stabla) dođe do zaključaka o output vrijednosti predmeta promatranja (prikazanom listovima stabla).

Na konkretnim podatcima promatrani su rezultati algoritma u kojima je najveća dopuštena dubina stable 2, 4, 6, odnosno 8 te su dobiveni sljedeći rezultati:

max_dept=2

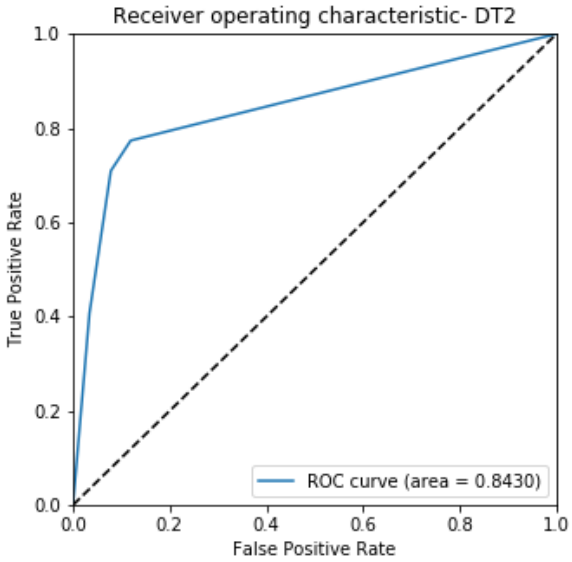
Dobivena točnost je

T=0. 90274405944236

a točnost na test setu je dobivena

Ttest= 0. 8948773974265598.

ROC krivulja za ovaj odabir maksimalne dubine stable prikazana je na slici 2.



Slika 2. : ROC krivulja, stable odlučivanja, max.dept=2.

Dobivene vrijednosti površine ispod ROC krivulje su

AUC=0.8430

AUCtest=0.8781 (za test set).

max_dept=4

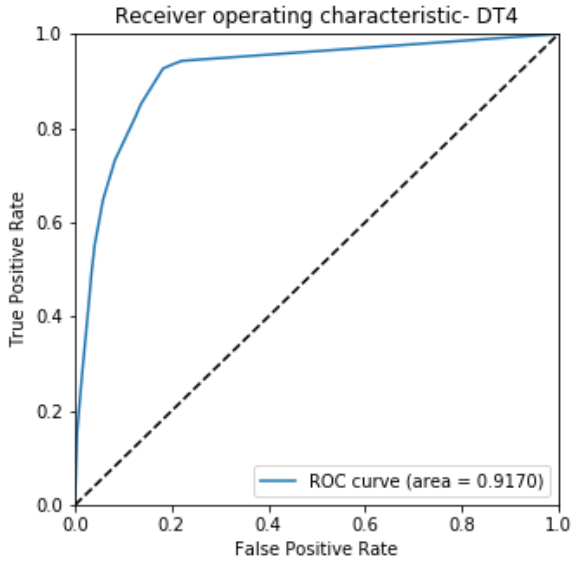
Dobivena točnost je

T=0. 9123078054881189

a točnost na test setu je dobivena

Ttest= 0. 9002184996358339.

ROC krivulja za ovaj odabir maksimalne dubine stable prikazana je na slici 3.



Slika 3. : ROC krivulja, stable odlučivanja, max.dept=4.

Dobivene vrijednosti površine ispod ROC krivulje su

AUC=0.9170

AUCtest=0.9240 (za test set).

max_dept=6

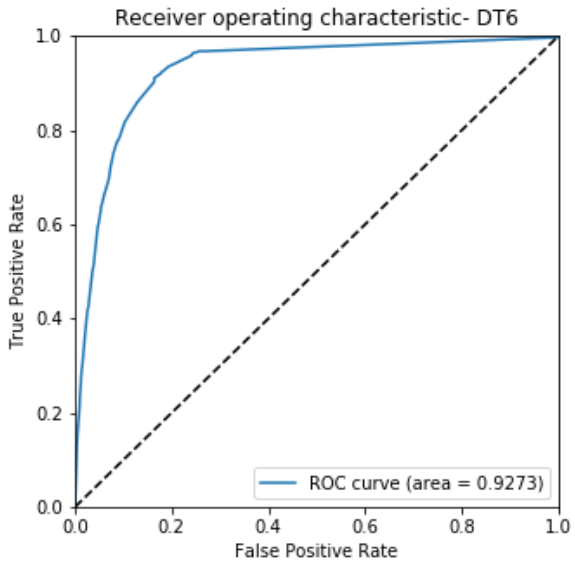
Dobivena točnost je

T=0. 9116456999926432

a točnost na test setu je dobivena

Ttest= 0. 8975479485311969.

ROC krivulja za ovaj odabir maksimalne dubine stable prikazana je na slici 4.



Slika 4. : ROC krivulja, stable odlučivanja, max.dept=6.

Dobivene vrijednosti površine ispod ROC krivulje su

AUC=0.9273

AUCtest=0.9323 (za test set).

max_dept=8

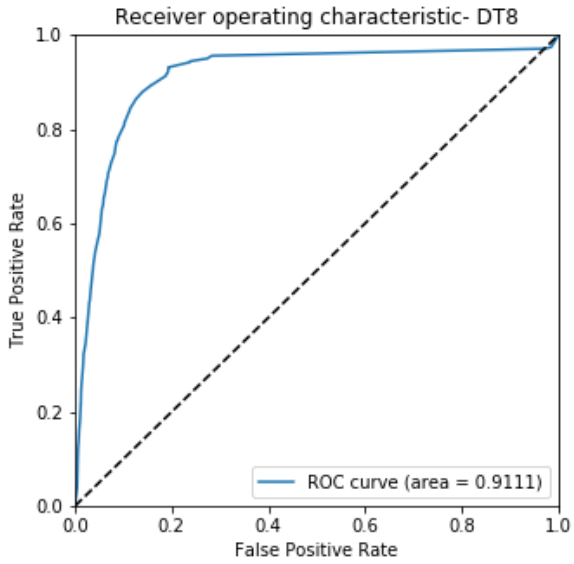
Dobivena točnost je

T=0. 9111307290517178

a točnost na test setu je dobivena

Ttest= 0.894634620053411.

ROC krivulja za ovaj odabir maksimalne dubine stable prikazana je na slici 5.



Slika 5. : ROC krivulja, stable odlučivanja, max.dept=8.

Dobivene vrijednosti površine ispod ROC krivulje su

AUC=0.9101

AUCtest=0.9133 (za test set).

Uspoređujući vrijednosti AUC možemo zaključiti da je od prethodna četiri modela

max_dept	AUC
2	0.8430
4	0.9170
6	0.9273
8	0.9101

najuspješniji onaj s najvećom vrijednošću AUC, a to je model stable odlučivanja s maksimalnom dubinom stable jednakom 6.
Za tako odabran model stable odlučivanja dobiva se da će biti ostvareno približno 93% uspješnih kontakata s klijentima, tj. onih koji će završiti klijentovim dugoročnim oročavanjem depozita.
Budući da je jedan od mogućih problema ove metode overfitting, sljedeća promatrana metoda je metoda slučajnih šuma, kojoj je cilj sprječavanje overfittinga.

Slučajne šume

Metoda slučajnih šuma (eng. random forest) generira ansambl stabla odlučivanja (engl. decision trees) i uprosječuje njihova predviđanja. Svako stablo je izgrađeno na slučajnom podskupu (s ponavljanjem) skupa za učenje i prilikom izgradnje svakog stabla se uvijek uzima slučajni podskup značajki za izgradnju svakog pojedinog čvora. Paranetar n_estimators regulira broj stabala dok parametar max_features regulira broj nasumično odabranih značajki koje se razmatraju prilikom izgradnje svakog čvora. Parametar max_depth regulira maksimalnu dubinu stabala u ansamblu.
U ovom radu je ova metoda korištena sa sljedećim kombinacijama značajki:

	Parametar		
	max_depth	n_estimators	max_features
Vrijednost parametra	5	2	1
	5	2	3
	5	4	1

Kao mjera uspješnosti promatrana je ROC-krivulja te su dobiveni sljedeći rezultati:

Slučaj 1.

max_depth	n_estimators	max_features
5	2	1

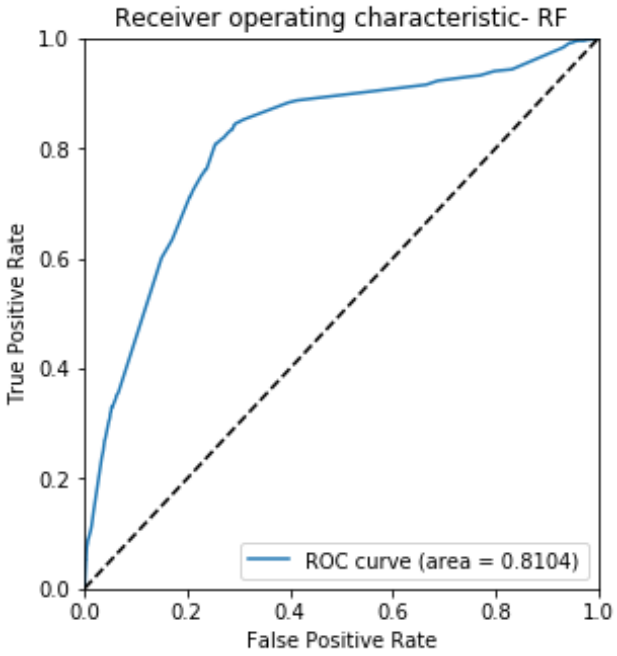
Dobivena točnost je

T=0.887515633046421

a točnost na test setu je dobivena

Ttest= 0.8909929594561787.

ROC krivulja za ovaj odabir parametara prikazana je na slici 2.



Slika 2. : ROC krivulja, metoda slučajnih šuma, slučaj 1.

Dobivene vrijednosti površine ispod ROC krivulje su

AUC=0.8104

AUCtest=0.8002 (za test set).

Slučaj 2.

max_depth	n_estimators	max_features
5	2	3

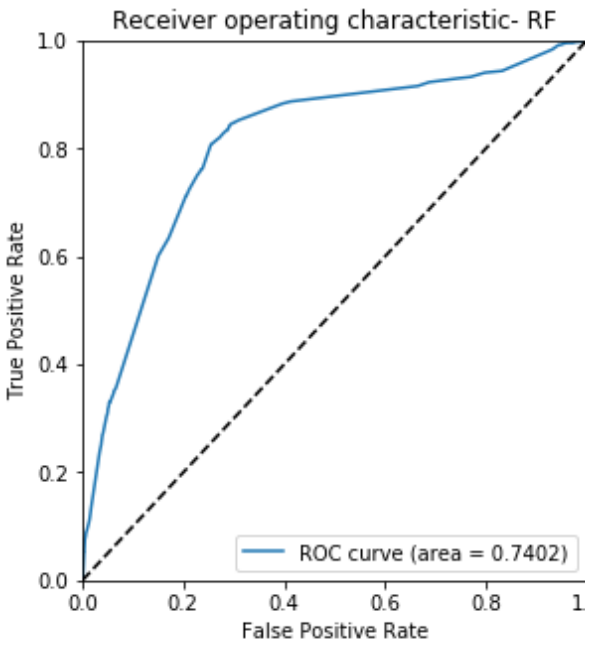
Dobivena točnost je

T= 0.896270139042154

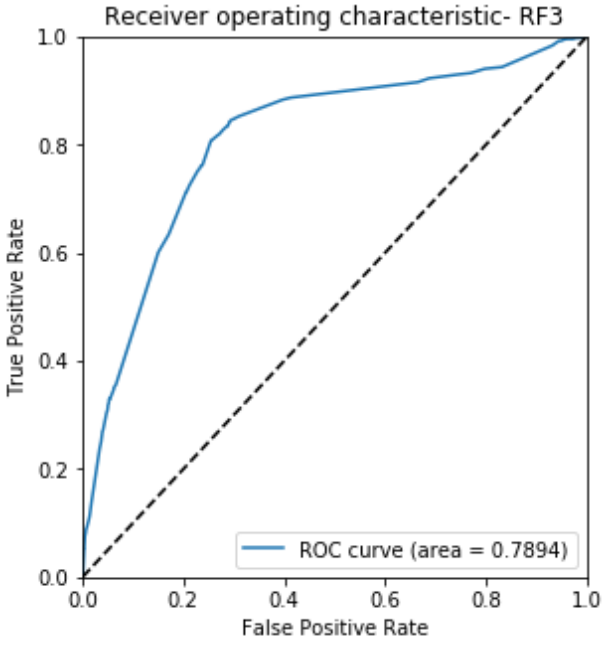
a točnost na test setu je dobivena

Ttest= 0. 8953629521728574.

ROC krivulja za ovaj odabir parametara prikazana je na slici 3.



Slika 3. : ROC krivulja, metoda slučajnih šuma, slučaj 2.



Slika 4. : ROC krivulja, metoda slučajnih šuma, slučaj 3.

Dobivene vrijednosti površine ispod ROC krivulje su

AUC=0.7402

AUCtest=0.7148 (za test set).

Slučaj 3.

max_depth	n_estimators	max_features
5	4	1

. Dobivena točnost je

T= 0.8870006621054954

a točnost na test setu je dobivena

Ttest= 0.890507404709881.

ROC krivulja za ovaj odabir parametara prikazana je na slici 4.

Dobivene vrijednosti površine ispod ROC krivulje su

AUC=0.7894

AUCtest=0.7877 (za test set).

Promatranjem dobivenih vrijednosti površina ispod ROC krivulje za odabrane slučajeve

slučaj	AUC
1	0.8104
2	0.7402
3	0.7894

može se zaključiti da je najbolji od ova tri klasifikatora onaj s najvećom vrijednošću AUC, a to je klasifikator sa sljedećim parametrima:

max_depth	n_estimators	max_features
5	2	1

Dakle, za najbolju kombinaciju parametara, ovom metodom će banka uspješno kontaktirati 81% klijenata spremnih dugoročno oročiti depozite.

Stroj s potpornim vektorima (SVM)

Stroj s potpornim vektorima (engl. support vector machine, SVM) rješava sljedeći optimizacijski problem:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i, \quad y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i, \quad \zeta_i \geq 0, i = 1$$

Gdje su {x1,…,xn} primjeri za treniranje a y∈{−1,1} ciljna značajka. Zbog efikasnosti često se zapravo rješava dualni problem:

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha, \quad y^T \alpha = 0, \quad 0 \leq \alpha_i \leq C, i = 1, \ldots, n$$

U ovom radu korišten je SVM implementiran u funkciji sklearn.svm.SVC koji interno koristi libsvm. Na podatcima je korištena funkcija sklearn.svm.SVC.decision_function() koja vraća udaljenost od decizijske hiperravnine. Potporni vektori su oni primjeri koji leže na udaljenost 1 od hiperravnine.

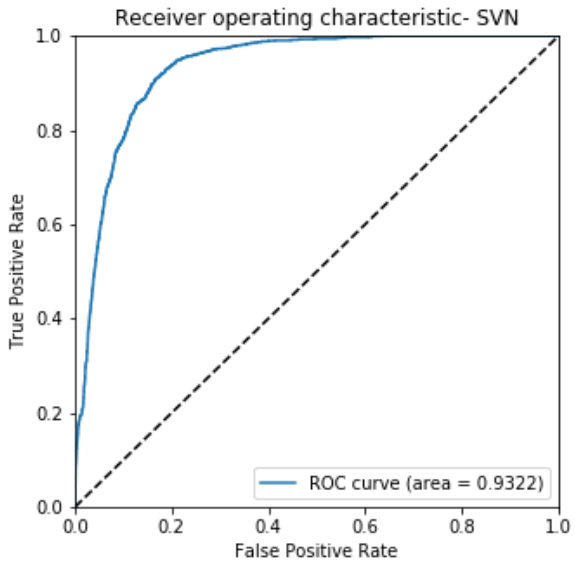
Dobivena točnost je

T= 0. 8968586772603546

a točnost na test setu je dobivena

Ttest= 0.9009468317552805.

ROC krivulja za ovaj odabir parametara prikazana je na slici 5.



Slika 5. : ROC krivulja, metoda potpornih vektora

Dobivene vrijednosti površine ispod ROC krivulje su

AUC=0.9322

AUCtest=0.9410 (za test set).

Dakle, sa stajališta banke, uspješno će biti kontaktirano 93% odabranih klijenata.

Zaključak

Osvrtom na sve promatrane algoritme strojnog učenja, cilj je odrediti najuspješniji od promatranih. U tablici 2. dan je popis svih promatranih algoritama te pripadne točnosti i AUC vrijednosti dobivenih ROC krivulja.

algoritam	točnost	AUC
naivni Bayes	0.11314647244905467	0.8304
stable odlučivanja	0. 9116456999926432	0.9273
slučajne šume	0.887515633046421	0.8104
SVM	0. 8968586772603546	0.9322

Tablica 2.: Usporedba korištenih algoritama

Kao kriterij mjere uspješnosti promatramo površinu ispod ROC krivulje. Algoritam je to uspješniji što je površina ispod krivulje veća. Prema tome, najuspješniji od korištenih algoritama je SVM, potom slijedi metoda stable odlučivanja, nešto manje uspješan je naivni Bayes, dok se najmanje uspješnim pokazala metoda slučajnih šuma.

Promatranjem do sada objavljenih radova, dobiveni rezultati u ovom radu slažu se sa do sada dobivenim rezultatima.

Od metoda korištenih i u ovom radu i u [1] te u [6], kao najbolja metoda pokazao se SVM.

Mogući budući nastavak istraživanja

Jedan od mogućih budućih pristupa je pokušaj odabira značajki koje bi doprinjele "točnijem", a ujedno i jednostavnijem modelu.

Kao na primijeru Random Forest modela (max_depth=5,n_estimators=2,max_features1) gdje se vidi koliko koja značajka pridonosi, tj je bitna u kreiranju modela:

```
Top 10 important variables:
              variable    weight
0             duration  0.229892
0        emp.var.rate  0.222231
0           month_oct  0.216377
0         job_student  0.072364
0    poutcome_success  0.068267
0          default_no  0.042923
0                 age  0.033795
0               pdays  0.022600
0           month_apr  0.020361
0  education_basic.9y  0.012350
```

Dakle, moguće je promatrati, primjerice samo one varijable čija je težina veća od 5%.

Na ovakav način promatrano bi bilo 5 značajki te output varijabla, što dovodi do jednostavnijih modela.

Literatura

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

[2] Choong, A.: Predictive Analytics in Marketing: A Practical Example from Retail Banking, Singapore Actuarial Society, 2017.

[3] Jiang, Y.: Using Logistic Regression Model to Predict the Success of Bank Telemarketing, International Journal on Data Science and Technology, Vol. 4, No. 1, 2018, pp. 35-41. doi: 10.11648/j.ijdst.20180401.15

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

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

[6] Wang, Y.: Building multiple machine learning models for success of bank telemarketing using scikit-learn, 2017.

[7] Yang, C.: Predicting Success of Bank Telemarketing with Classification Trees and Logistic Regression,2016., diplomski rad

[8] Cultivating the Customer Relationship in Banking, A. T. Kearney, izvještaj, 2017.

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

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

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

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

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

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