Предвиђање исхода маркетиншке кампање банке

Милош Младеновић
Факултет техничких наука
Универзитет у Новом Саду
milosml@outlook.com

***Сажетак* — Маркетиншке кампање банака и пословних институција су важно средство за стицање нових клијената и испитивање стања на тржишту, с обзиром на то да већи економски фактори најдужи век деловања имају на обично становништво, које броји и највећи део клијената једне банке. Добра стратегија планирања и одабира потенцијалних корисника за контактирање приликом маркетиншке кампање је од велике важности за успех исте. Технике анализе података као што су оне у Data mining-у могу у многоме да помогну банкама и пословним институцијама да лакше пронађу клијенте и понуде / продају им свој производ или услугу, тиме побољшавајући свој финансијски успех. Идеја овог пројекта је истраживање и упоређивање перформанси алгоритама за класификацију на скупу података који је прикупљан у периоду од 2008-2013. године од стране једне португалске банке, а који се односи на њихову телефонску маркетиншку кампању у којој покушавају да убеде потенцијалне клијенте да заложе новац у њиховој банци. Детаљно истраживање података и извлачење могућих закључака из њих је такође спроведено. Коришћени су следећи алгоритми машинског учења: *SVM (support vector machines), Gaussian Naïve Bayes, Decision Trees, Random Forest, Bagging, Boosting, Logistic Regression*. Због лоше избалансираности циљних класа, упоређене су перформансе наведених алгоритама након примене две технике узорковања оригиналног скупа података приликом тренирања модела: „SMOTE" и "majority class undersampling". Најбоље резултате на тест скупу је показала *Логистичка регресија* уз примену технике *SMOTE*, и то вредност *AU ROC* од 0.73 и однос *FN/TP* – 343/576, чиме је потврђено да је модел валидан и од могуће користи менаџерима маркетиншких кампањи банака.**

***Кључне речи* — истраживање података, класификација, машинско учење, статистика, банка, маркетинг**

I. УВОД

Напредак комуникационих технологија донео је и нове могућности у остваривању маркетиншких циљева пословних организација и компанија, а један од најважнијих облика међусобне интеракције пословних институција и грађана постале су телефонске маркетиншке кампање (телемаркетиншке кампање скраћено, у наставку). Телемаркетиншке кампање су један од главних извора нових клијената пословних организација, а посебно банака, као и добар показатељ тренутног јавног мњења које влада поводом одређених питања која су од интереса банкама и утичу великим делом на изворе њихових прихода и кретања на тржиштима. Телемаркетинг је заправо врста „директног маркетинга" – термина који је први пут био предложен 1967. године од стране Лестара Вундермана, који се због тога сматра „оцем" директног маркетинга. [1]

Банке на овај начин, као и кроз само пословање са становништвом скупљају велику количину података који касније могу да буду коришћени да се оствари индивидуални директни контакт са потенцијалним клијентима и са добром сигурношћу предвиди успех тог контакта у смислу да ће контактирана особа постати клијент банке у неком облику. Директни контакт се најчешће остварује телефонским позивом, било фиксним или мобилним, електронском поштом, обичним писмом или личном посетом. [2]

До скора, главно средство маркетинга које су банке и пословне институције користиле био је масовни маркетинг, путем телевизије или штампаних медија, међутим, развој директног маркетинга омогућио је да пословође имају директан увид у успешност водиоца маркетиншке кампање, али и да савремена достигнућа у области истраживања и анализе података употребе као средства за побољшање пословања и лакше стицање нових клијената.

Системи подршке одлучивању (енг. *DSS – Decision support systems*) користе информационе технологије за помоћ и подршку одлукама које доносе менаџерски тимови. Постоји неколико *DSS* области, као што су лични и интелигентни *DSS*. Лични *DSS* су повезани са системима ситних размера, који служе за помоћ при одлучивању једном менаџеру, док интелигентни *DSS* користе технике вештачке интелигенције да помогну у доношењу одлука. [3]

Главни концепт повезан са *DSS*-ом јесте Истраживање и анализа података (енг. *Data mining*), где алгоритми машинског учења употребљавају за полу-аутоматско извлачење скривеног и предиктивног знања из података. Основну улогу у решавању овакве врсте проблема имају класификациони алгоритми, као што су *Стабла одлучивања* (енг. *Decision Trees*), *Случајне шуме* (енг. *Random Forests*), Логистичка *регресија* (енг. *Logistic regression*), *Наивни Бајес* (енг. *Naïve Bayes*), *Машине потпорних вектора* (енг. *Support Vector Machines*) итд. Такође, детаљна експлоративна анализа података кроз цртање графикона и упоређивање зависности исхода неке предвиђене хипотезе од доступних података могу довести до значајних открића.

Проблем којим се бави овај рад јесте управо креирање модела за предвиђање исхода телемаркетиншке кампање, коју је спровела једна португалска банка и откривање зависности прихватања или одбијања за залог депозита од фактора као што су године, пол, примања клијента, тренутно стање у економији, упорност менаџера у

инсистирању на позивима итд. Скуп података који је у ту сврху коришћен преузет је са UCI репозиторијума, а односи се на податке које је португалска банка скупљала у периоду од 2008-2013 године, за које време су водили телемаркетиншку кампању где су настојали да убеде потенцијалне кориснике да ороче новац на одређено време у њиховој банци. Идеја је била и да се открију параметри који највише утичу на исход маркетиншке кампање, као и њихов тип - лични (зависан од клијента) или општи (тренутна економска ситуација) како би се пронашли неки шаблони и будуће кампање успешније планирале и изводиле. Показало се да је најбоље перформансе у смислу успешности предвиђања на тест скупу имао модел изграђен помоћу *SVM* алгоритма за класификацију у случају када су прикупљени подаци семпловани техником *SMOTE* која ће касније бити детаљније објашњена. Површина под кривом *ROC* код овог алгоритма износила је 0.75, док је *F1* мера била 0.91.

У наставку рада ће детаљније бити објашњени различити аспекти решаваног проблема и самог решења. У поглављу **II** је направљен кратак преглед пронађених радова који се баве истом или сличном проблематиком. За сваки рад је детаљније образложена методологија којом је проблем решаван као и добијени резултати. Поглавље **III** садржи детаљнији опис скупа података, као и главне карактеристике скупа уз наведене мане које су морале касније у процесу креирања модела бити исправљене да би се добио бољи модел. Поглавље **IV** садржи статистичке анализе спроведене над скупом података, односно најважније закључке до којих се дошло експлоративном анализом података. Поглавље **V** описује процес креирања модела за предикцију, испробане алгоритме, методе узорковања података приликом креирања модела као и дискусију најефикаснијих модела и резултата који су њиховом применом били добијени. Напослетку, поглавље **VI** садржи дискусију на тему добијених резултата, сумаризацију рада, правце будућег развоја као и могуће предлоге за даље унапређење модела или самог процеса прикупљања сличних типова података.

II. Претходна решења

У наставку овог поглавља биће наведена два најрелевантнија решења везана за овај проблем, по мишљењу аутора, која су пронађена на интернету:

A. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems [4]

Тема овог рада је била формирање модела који предвиђа успех маркетиншке кампање за понуду дугорочне штедње клијентима којима је упућен телефонски позив од стране португалске банке, на основу личних података о потенцијалним клијентима, као и тренутној економској ситуацији у земљи.

Иницијални сет од 150 параметара, аутори су након експлоративне анализе и редукције димензионалности свели на 22 параметра, укључујући и оне личне о клијентима као што су: пол, године, ниво образовања, брачни статус, кредитни и хипотекарни статус итд. али и општије податке: трајање позива упућеног клијенту од стране маркетинг менаџера, број остварених позива, претходни исход позива итд. Такође, неки од параметара који су се показали интересантним и укључени у креирање модела су неки од оних социјалних и економских на нивоу државе: тренутни број запослених, индекс о поверењу потрошача, euribor3 месечни ниво итд.

Алгоритми који су у овом раду били коришћени су Логистичка регресија, Стабла одлуке, Неуронске мреже и Машине потпорних вектора. Модел који је развијен помоћу Неуронских мрежа се показао посебно добрим за предвиђање успеха телемаркетиншке кампање и може бити од велике помоћи менаџерима оваквих кампања. Применом метода за анализу закључака на моделу креираном неуронским мрежама дошло се до неких изненађујућих резултата о параметрима који највише утичу на исход кампање, као што је euribor3 ниво, смер позива (од банке ка клијенту или обрнуто) и искуство агента који обавља позив. Потенцијални непредвиђени општи фактор који би могао да утиче на ове резултате јесте велика светска економска криза која је достигла 2008-2009. године, па би податке из овог периода можда требало издвојити за посебан модел, што овде није учињено. Такође, у раду није напоменуто како је решен проблем лоше избалансираности класа, у смислу да је тек око 10% скупа садржало као циљну класу „*Yes*“.

B. Bank direct marketing analysis of data mining techniques [5]

У овом раду главни фокус аутора је био стављен на проналажење најважнијих атрибута који утичу на успех маркетиншке кампање.

Скуп података потребан за овај рад је преузет са Machine Learning репозиторијума UCI (University of California at Irvine), а подаци који су ту били присутни скупљени су раније од португалских истраживача S. Moro, R. Laureano и P. Cortez. Подаци су везани за маркетиншку кампању португалске банке, базирану на телефонским позивима, чији је циљ био да убеди потенцијалне клијенте да оставе депозит у банци. Скуп података који је коришћен је онај скраћени, односно састоји се од 17 атрибута: старост, образовање, контакт, месец, претходни резултат позива, трајање позива, трајање претходног позива, кредит, хипотека, брачни статус итд.

Алгоритми које је аутор овог рада користио били су *Tree augmented Naïve Bayes (TAN),* Логистичка регресија, Рос-Квинланово стабло одлуке (C5.0) и вишеслојна неуронска мрежа.

Циљ овог рада је био да се идентификују параметри који највише утичу на исход маркетиншке кампање. Од тренираних модела најбоље се показао онај који је настао применом C5.0 алгоритма, остваривши резултате прецизности од чак 93% на тренинг скупу и 90% на тест скупу. Као атрибут од кога највише зависи исход кампање је идентификовано „трајање позива“ упућеног од менаџера кампање клијенту или обрнуто за C5.0, неуронску мрежу и

логистичку регресију, а код TAN алгоритма то је био атрибут „старост клијента“. Овај рад је остварио добре резултате приликом предвиђања исхода, међутим, није обухватио све атрибуте којима је тек касније скуп података проширен, као што су они општи социјално-економски у држави у том тренутку, а који можда могу да баце другачије светло на исход предвиђања оваквих кампања. Такође, још један проблем на који аутор није обратио пажњу или бар није наглио у раду јесте лоша избалансираност класа.

III. Скуп података

Скуп података који је у овом раду коришћен скуп је доступан на интернету, конкретно на UCI репозиторијуму: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

садржи податке о исходима маркетиншке кампање неименоване португалске банке у периоду од маја 2008. до новембра 2010. године. Подаци су били доступни у *.csv* датотеци. Цео скуп је величине 41190 инстанци, циљни атрибут у скупу је ***y*** – који представља резултат кампање, да ли је корисник пристао на штедњу или не. У *табели 1* приказан је скуп атрибута са детаљним описом шта сваки атрибут значи и које вредности може да има. Класе нису једнако заступљене, односно број корисника који су одбили да штеде је знатно већи у односу на број оних који су пристали и то у односу: 35400:5790. Атрибути су

Број	Назив атрибута	Тип	Вредности	Опис
1.	age	нумерички		
2.	job	номинални	‘admin’, ‘blue-collar’, ‘enterpreneur’, ‘housemaid’, ‘management’, ‘retired’, ‘self-employed’, ‘services’, ‘student’, ‘technician’, ‘unemployed’, ‘unknown’,	Тип посла којим се бави испитаник
3.	marital	номинални	‘divorced’, ‘married’, ‘single’, ‘unknown’	Брачни статус испитаника
4.	education	номинални	‘basic.4y’, ‘basic.6y’, ‘high.school’, ‘illiterate’, ‘professional,course’, ‘university.degree’	Ниво стручне спреме испитаника
5.	default	номинални		
6.	housing	номинални	‘no’, ‘yes’, ‘unknown’	Да ли испитаник има хипотеку на кућу
7.	loan	номинални	‘no’, ‘yes’, ‘unknown’	Да ли има кредитну позајмицу од банке
8.	contact	номинални	‘cellular’, ‘telephone’	Тип оствареног телефонског контакта
9.	month	номинални	‘january’, ‘february’..	Месец у ком је остварен последњи контакт
10.	day_of_week	номинални	‘Monday’, ‘Tuesday’…	Дан телефонског позива
11.	duration	нумерички		Трајање телефонског позива кад је сазнат исход кампање
12.	campaign	нумерички		Број контаката истог испитаника
13.	pdays	нумерички		Број дана протеклих од позива истом клијенту
14.	previous	нумерички		
15.	poutcome	номинални	‘failure’, ‘success’, ‘nonexistant’	Исход претходне кампање
16.	emp.var.rate	нумерички		Стопа варирања запослености
17.	cons.price.idx	нумерички		Индекс потрошачких цена
18.	cons.conf.idx	нумерички		Индекс поверења потрошача
19.	euribor3m	нумерички		Euribor3m месечни ниво
20.	nr.employed	нумерички		Број запослених - квартално
21.	y (outcome)	номинални	‘yes’, ‘no’	Исход кампање

Табела 1: Атрибути скупа података

Подељени у три групе: прва група су они који се односе на личне информације корисника: 1-7. Другу групу чине атрибути који се односе на на маркетиншку кампању банке у односу на истог корисника: 8-15, док трећој групи припадају атрибути који се тичу тренутног социјално-економског стања у земљи: 16-21.

На сајту UCI репозиторијума доступна су два скупа - један са редукованим бројем атрибута и други са оригиналним скупом, као што је наведено у табели.

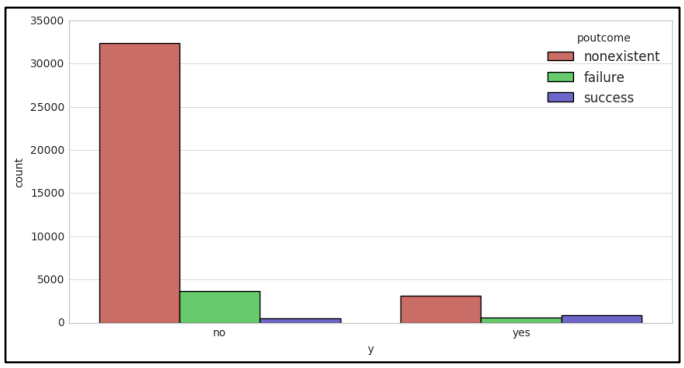
Редуковани скуп атрибута не садржи последњу групу – социјално-економске атрибуте, 16-21, и то је онај коришћен у раду [5]. Такође, скуп података доступан на сајту је морао да буде обрађен на више начина, а први се односио на атрибут *месец*, где су били наведени редом месеци како је скупљан скуп података, без наведених година, што је морало да буде преправљено, да буде конзистентно са временским током.

IV. Експлоративна анализа

У овом поглављу су наведене главне статистичке и експлоративне анализе које су спроведене над скупом података, као и закључци до којих се њиховом применом дошло.

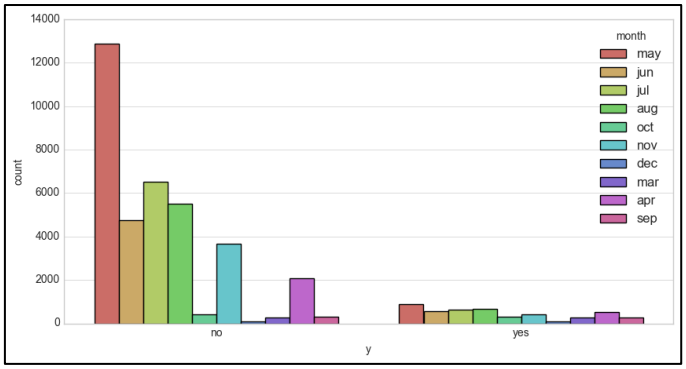
Провером скупа података по свим атрибутима, закључено је да није било недостајућих података, као ни очигледних аутлајера, насталих због погрешног уноса, људске грешке и сл. што значи да су сви подаци били валидни и спремни за анализу. У наставку следе неки закључци из процеса експлоративне анализе.

- Са *дијаграма 1* можемо видети да већина контактираних корисника претходно није учествовала у кампањи, али да више од 50% оних који су учествовали и претходно прихватили да заложе новац, поново ће то урадити



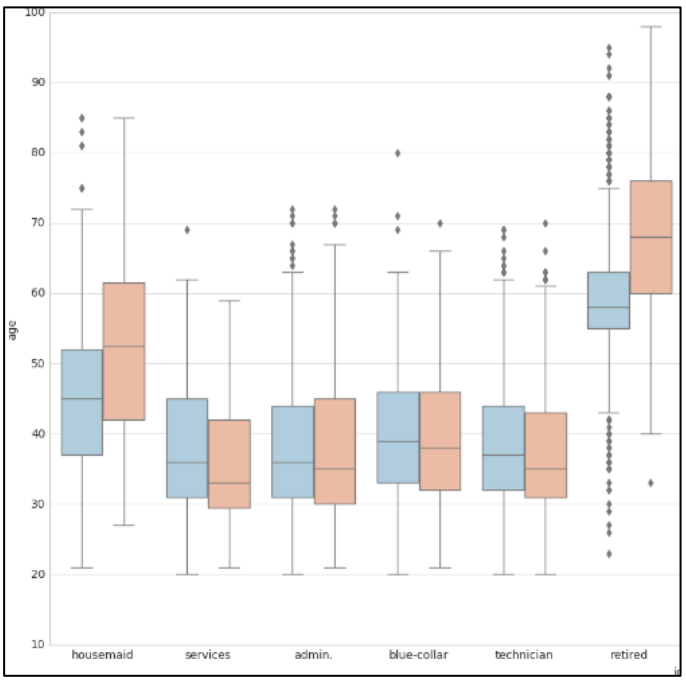
Дијаграм 1: Исход претходне кампање - исход тренутне

- На *дијаграму 2* уочљиво је да у месецима октобар и март, иако не постоји велики узорак контактираних корисника, вероватноћа да ће они позвани изабрати да заложе новац је скоро 50%.



Дијаграм 2: Број контактираних по месецима - исход

- *Дијаграм 3* представља кутијасти дијаграм просечне старости контактираних људи у односу на њихово занимање, као и исход кампање. Занимљив закључак до којег и здравом логиком можемо доћи, а који потврђује и анализа јесте да је већа вероватноћа да ће пензионисана особа постати клијент банке што је старија, што се може видети по просечној старости у „retired" делу за плави (одбијено) и црвени (прихваћено) дијаграм. Ово се дешава из простог разлога што старији пензионери имају више уштеђеног новца од плата и пензија, који могу да оставе у банци, а који њима тренутно није неопходан за трошак, али може

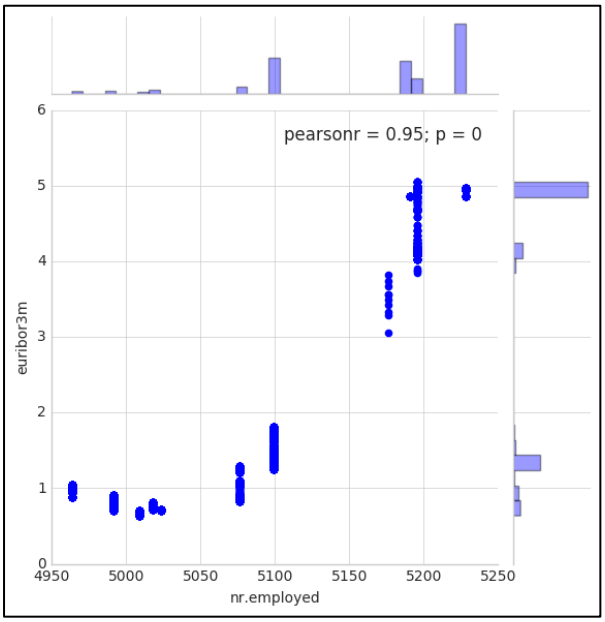


Дијаграм 3: Просечна старост људи у односу на занимање и исход кампање

касније , са каматом, послужити потомцима.

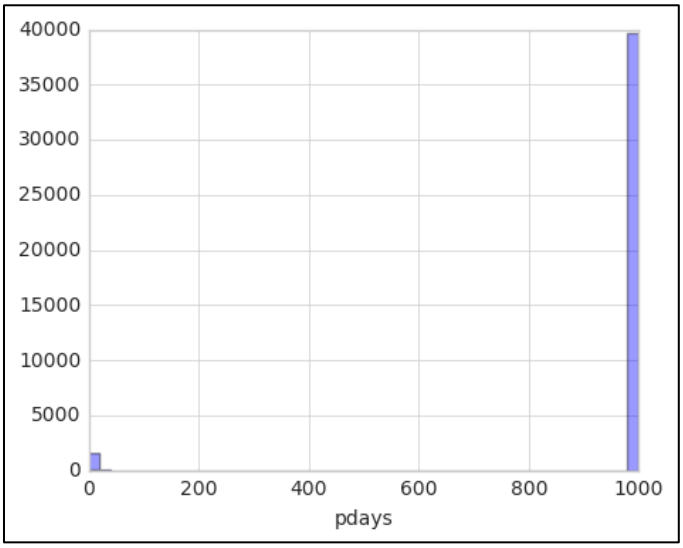
- На *дијаграму 4* може се видети корелација између два атрибута – *euribor3m* и броја запослених, где се јасно види да је коефицијент Пеарсонове

корелације 0.95, што значи да су атрибути јако међусобно корелирани, па је један од њих у каснијем креирању модела за предикцију и био уклоњен (број запослених).



Дијаграм 4: Корелација између атрибута euribor3m и бр. запослених

- Још један пар променљивих које су високо корелиране (0.97 коефицијент корелације) , а чији дијаграм неће бити приказан су *euribor3m* и *emp.var.rate* , тако да је у каснијем креирању модела друга променљива изузета.
- На *дијаграму 5* може се видети и дистрибуција променљиве *pdays* где је очигледно да она може врло непогодно утицати на модел јер има само распон од две „доминантне“ вредности – 999 и 0, односно вредности које су око њих, због тога што је приликом формирања скупа података у поље „број дана протеклих од последњег контакта“ – *pdays* уношена вредност 999, уколико није ниједном остварен контакт, тако да модели машинског учења због својих математичких операција могу да буду ометени овом променљивом са оваквом дистрибуцијом, тако да је она из финалног модела отклоњена.

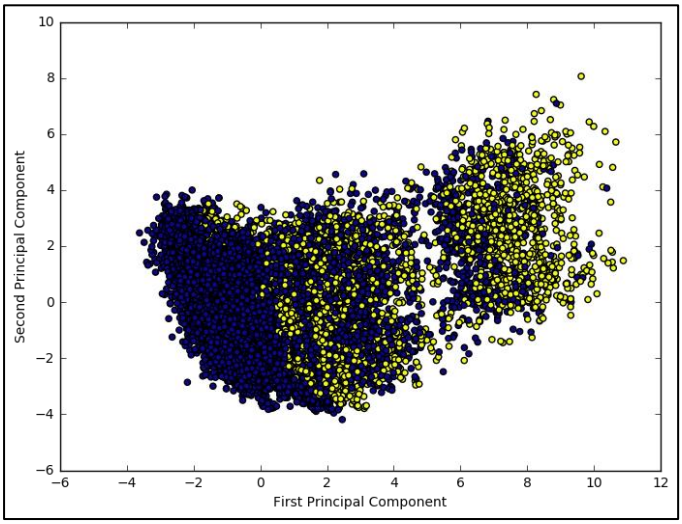


Дијаграм 5: Дистрибуција променљиве pdays

Генерално, статистичка односно експлоративна анализа података омогућила је увид у неке скривене информације о скупу података, које би могле бити од користи менаџерима маркетиншке кампање у будућим подухватима, које делове популације да циљају, чак и без самог предиктивног модела. Наравно, сем тога, анализа је помогла и око селекције атрибута за будуће предиктивне моделе, јер је пружила информације о високо корелираним атрибутима, које је потребно уклонити пре креирања модела. Још један облик анализе података пре самог креирања модела јесте и *Анализа главних компоненти* (енг. *Principal Component Analysis*), што је модел ненадгледаног учења, где се заправо скуп из оригиналних димензија пројектује на нове димензије, које представљају векторе највеће варијабилности података. У следећем одељку биће више речи о њој, као и њеним резултатима након примене на овом скупу података.

A. PCA (Principal Component Analysis)

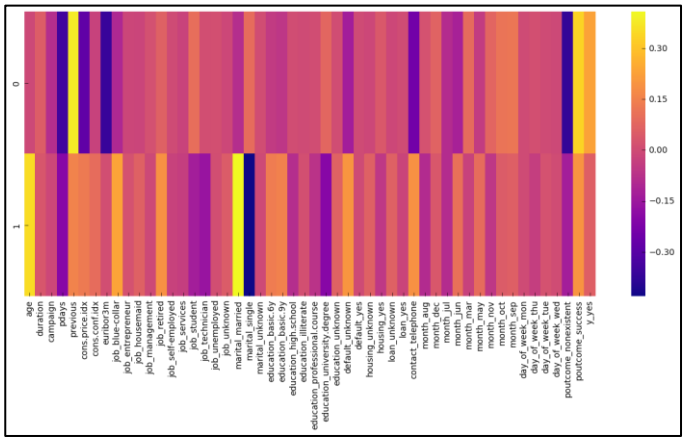
Анализа главних компоненти (у наставку – PCA) представља технику за редукцију димензионалности података, тако што се из тренутних димензија (атрибута) примери из скупа података пројектују на нове димензије, које представљају димензије највеће варијабилности података. Редукција димензионалности се касније врши тако што се узме првих неколико ових компоненти, пошто



Дијаграм 6: PCA са две компоненте

оне објашњавају највећи део варијабилности. Самим тим, могуће је оригинални скуп од неколико десетина атрибута свести на само два или три, ради лакшег приказивања и уочавања неких правилности и особина података. На *дијаграму 6* приказан је оригинални скуп података након примењене *PCA* и свођења на две компоненте. Плавом бојом су обележене особе које су одбиле да буду клијенти банке, а жутом оне које су прихватиле. Могуће је уочити да „плави“ примери углавном имају већу кохезију, али да су „жути“ већином помешани са „плавима“, што значи да немају нешто што их стриктно издваја од других, што их одликује.

Још једна занимљива и корисна ствар која се применом *PCA* редукције димензионалности може добити јесу коефицијенти који објашњавају варијабилност по свакој оси која учествује у формирању неке од *главних компоненти*, а који стоје уз сваки оригинални атрибут из скупа података. Ови коефицијенти нам говоре колико дати атрибут утиче на формирање *главне компоненте*, па уколико неки атрибути имају слаб или никакав утицај на првих неколико главних компоненти (које објашњавају преко 90% варијабилности), онда они обично неће бити ни превише значајни у процесу креирања модела, па их је могуће уклонити. Пример како је *PCA* коришћен за селекцију атрибута приказан је на *дијаграму* 7, са којег је прегледом коефицијената уз сваки атрибут утврђено да атрибути *'job_blue-collar', 'job_self-employed', 'job_services', 'job_unemployed'* и *'day_of_week_mon'* најмање утичу на варијабилност, па су код креирања модела уклоњени, што је довело до побољшања перформанси.



Дијаграм 7: Коефицијенти уз PCA за прву и другу главну компоненту

Након овакве детаљне експлоративне анализе и процеса селекције атрибута путем редукције димензионалности, уследило је креирање модела за предикцију уз примену разних алгоритама за класификацију, чији ће резултати и процес бити описани у следећем поглављу.

V. МОДЕЛИ ЗА ПРЕДИКЦИЈУ ИСХОДА КАМПАЊЕ

Проблем који је у овом раду било потребно решити је класификациони проблем – односно предвиђање да ли ће контактиране особе постати клијенти банке или не. У ту сврху, коришћени модели за истраживање података базирани на алгоритмима за класификацију. Алгоритми који су у раду били испробани су: *SVM (support vector machines), Gaussian Naïve Bayes, Decision Trees, Random Forest, Bagging, Boosting, Logistic Regression*. Параметри модела за предикцију били су код свих алгоритама оптимизовани на валидационом скупу и касније су модели тестирани на посебном, дотад нетакнутом тест скупу. Поред оптимизације параметара модела, проблем на који је требало обратити пажњу је велика неизбалансираност циљних класа - односно број корисника који су одбили да штеде је знатно већи у односу на број оних који су пристали и то у односу: 35400:5790. Решење тог проблема је потражено на два различита начина: ређим узорковањем већинске класе приликом креирања тренинг скупа и применом *SMOTE* технике.

У наставку биће дат кратак опис рада *SMOTE* технике.

Пре него што буде дат опис креирања модела и постигнутих резултата, треба додати и да пошто су алгоритми за класификацију, односно модели за предвиђање исхода кампање креирани у *python*-у, коришћена је библиотека *scikit-learn* и код ње улазни атрибути за скуп података над којим се тренира модел и тестира морају да буду нумеричке вредности, тако да су номинални атрибути превођени у нумеричке путем *one-hot-encoding*-а. Ова техника ради тако што се за сваку могућу вредност номиналног атрибута креира нова колона атрибута, где се за сваки слог у скупу података уписује вредност 0 или 1 у зависности да ли у оригиналном скупу тај слог има вредност именоване колоне или не.

Такође, пре креирања сваког модела, уклоњени су атрибути: *‘duration’* – зато што је то атрибут који говори о трајању последњег позива где је заправо сазнат исход кампање, тако да нема смисла користити га у предиктивним моделима; *‘nr.employed’* и *‘emp.var.rate’* јер су у великој корелацији са *‘euribor3m’* – који је задржан.

A. *SMOTE (synthetic minority oversampling technique) [6]*

SMOTE техника се користи када имамо лошу избалансираност циљних класа и подразумева да се нови примери из циљне - „мањинске“ класе синтетички креирају тако што се насумично узме један од суседа сваке постојеће инстанце, па се вектору атрибута узете инстанце дода разлика вектора атрибута суседа и узете инстанце помножена неким насумичним бројем од 0 до 1. На овај начин, SMOTE техника ублажава *overfitting* који се јавља приликом примене „наивне“ класичне технике већинског семпловања мањинске класе, односно креирања поновљених примера који постоје у тренинг скупу, како би се изједначио. Горе описани начин је један од начина имплементације ове технике, а она постоји имплементирана

као *python* библиотека у пакету *imblearn* и коришћен је у овом раду.

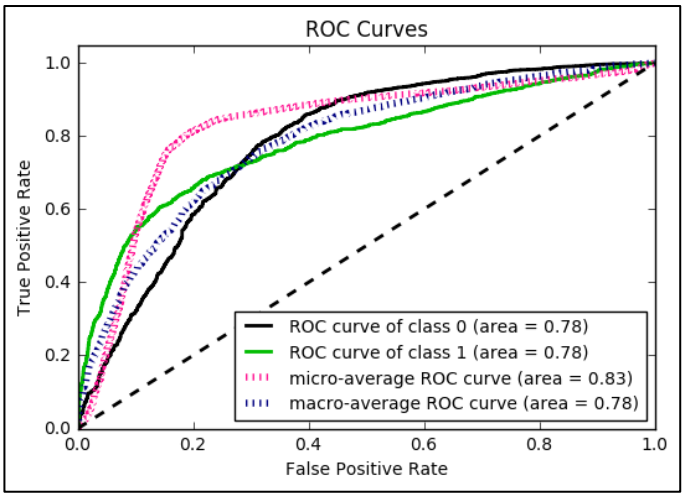
B. Модели тренирани на скупу података креираним ређим узорковањем већинске класе

Ово је једна од техника која се користи за решење проблема лоше избалансираности класа у предиктивним моделима и показала је прилично добре резултате, упоредиве са оним добијеним применом *SMOTE* технике. У наредној *табели 1* налазе се резултати постигнути на тест скупу применом сваког од алгоритама раније наведених у односу на 3 мере – *f1 меру, AU ROC (area under receiver operator characteristic)* и *број FN примера,* зато што је то најважнија ставка у матрици конфузије за наш проблем. Овакав је случај зато што *FN (false negative)* примери представљају заправо број „изгубљених" клијената банке због грешака модела, односно број који желимо да минимизујемо.

	f1 score	FN	AU ROC
Decision Trees (DT)	0.77	371	0.65
AdaBoost (DT)	0.83	414	0.69
Gradient Boosting	0.87	410	0.73
Bagging (DT)	0.85	415	0.71
Random Forest	0.85	396	0.72
Logistic Regression	0.83	**348**	**0.73**
SVM	0.87	412	0.73
K-nearest neighbors	0.87	511	0.69

Табела 2: Резултати класификатора коришћењем ређег узорковања већинске класе

Посматрајући резултате у табели изнад, могуће је извући различите закључке у погледу најбољег класификатора, у зависности од тога који критеријум одаберемо као најбитнији, да ли *f1* меру, *FN* или *AU ROC*. Ако *AU ROC* меру узмемо као најбитнију, као што је урађено у [4], имамо 3 кандидата за најбољи класификатор због једнаких резултата, па уколико од њих одаберемо најбољи по броју *FN* примера, добијамо да је *Логистичка регресија* заправо дала најбоље резултате и да је модел базиран на њој најбољи модел за предикцију исхода телемаркетиншке кампање. Изглед *ROC* криве овог модела примењеног на тест скупу дат је на *дијаграму 8*.



Дијаграм 8: ROC (receiver operating characteristic) крива за Логистичку регресију

C. Модели тренирани на скупу података креираним применом SMOTE технике на мањинској класи

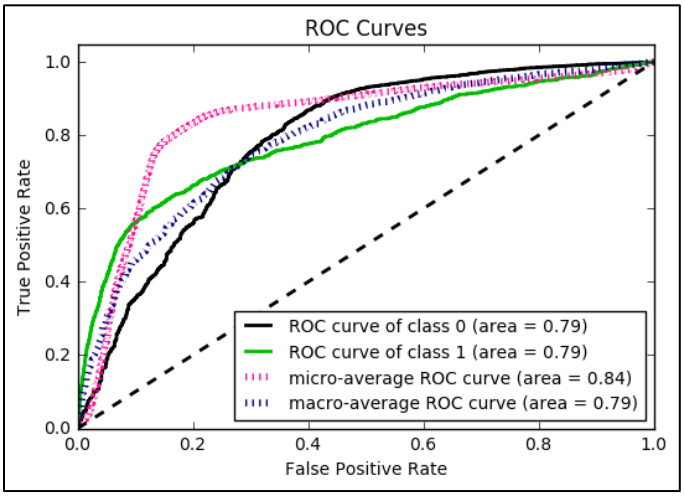
Техника SMOTE је раније описана, а принцип креирања модела, сем саме разлике у начину узорковања оригиналног скупа података, је истоветан оном описаном у претходном делу, па су у наставку приказани само резултати тренираних модела. Ваља још напоменути да је код модела базираног на SVM-у једино било, након примене SMOTE применити и *undersampling* целокупног тренинг скупа, због великог броја примера који су применом прве технике настали (> 70000), што је било превише рачунски захтевно за систем који је био на располагању. У *табели 3* налазе се резултати постигнути на тест скупу применом истог низа алгоритама описаним у претходном поглављу.

	f1 score	FN	AU ROC
Decision Trees (DT)	0.84	605	0.62
AdaBoost (DT)	0.87	657	0.62
Gradient Boosting	0.88	591	0.65
Bagging (DT)	0.88	605	0.64
Random Forest	0.88	638	0.63
Logistic Regression	0.84	**343**	**0.73**
SVM	0.88	475	0.70
K-nearest neighbors	0.85	583	0.63
Gaussian Naïve Bayes	0.86	506	0.67

Табела 3: Резултати класификатора коришћењем SMOTE технике

Поређењем свих резултата класификатора у табели изнад, могуће је закључити да је још једном модел базиран на

Логистичкој регресији дао најбоље резултате у погледу *AU ROC и FN*. На *дијаграму 9* испод биће приказан изглед *ROC* криве овог модела.



Дијаграм 9: ROC (receiver operating characteristic) крива за Логистичку регресију

VI. Закључак

У овом поглављу је дата дискусија на тему добијених резултата, праваца будућег развоја, могућих побољшања модела и идеја за унапређење пословања.

A. *Дискусија*

Анализом свих модела креираних помоћу две претходно описане технике, може се видети да је најбољи модел онај креиран помоћу Логистичке регресије и технике *SMOTE*, где је постигнут резултат од *AU ROC 0.73* и укупан број *FN* примера 348 од 919 примера мањинске класе. Ово је нешто слабији резултат од оног добијеног у [4], где је коришћена неуронска мрежа, а код ког је *AU ROC* износила 0.80, уз напомену да су истраживачи који су радили на том пројекту имали на располагању и сет од још више десетина атрибута, па су имали прилику да лакше оптимизују параметре за своје предиктивне моделе. Као што се из табела наведених у поглављу V може видети, Стабла одлучивања су дала релативно лоше резултате у односу на остале моделе, што је помало изненађујуће, због самог начина њиховог функционисања и преферирања великог броја номиналних променљивих, које су у овом скупу података доминирале. Због тога, као и због системских ограничења система на коме су обучавани и тестирани модели, стабла одлуке нису ни била визуализована.

B. *Сумаризација рада*

Оптимизација циљних група за контактирање приликом вођења телемаркетиншких кампања је од суштинског значаја за успех банака, под све већим притиском да се у банкарском свету повећа профит, а смање трошкови. Коришћење предиктивних модела и техника анализе података у многоме може да олакша посао менаџерима кампања.

Проблем којим се бави овај рад је предвиђање исхода телефонске маркетиншке кампање банке, која контактирањем особа о којима већ поседује неке информације у бази података из ранијих кампања или на други начин, покушава да их убеди да заложе новац код ње. Овај исход зависи од многих фактора као што су године, пол, примања клијента, тренутно стање у економији и упорност менаџера у инсистирању на позивима, општа економска ситуација у земљи итд. У овом раду је настојано да се овакви најбитнији фактори открију и употребе за креирање предиктивног (класификационог) модела, а да се они ирелевантни атрибути уклоне.

Први корак ка креирању модела за предикцију је био детаљна експлоративна анализа скупа података, уочавање међусобних зависности променљивих, уклањање аномалија и грешака из скупа и уклањање променљивих које нису релевантне за проблем и/или резултат. Следећи корак се бавио решавањем проблема лоше избалансираности класа у скупу података, који је решен на два различита начина и понуђени су резултати за оба. Уследило је тренирање модела, а затим евалуација и одабир оног најбољег по селектованим критеријумима, односно мерама евалуације.

C. *Правци будућег развоја*

Скуп података понуђен на UCI репозиторијуму, иако прилично детаљан и обиман, ипак не пружа све слогове који су били скупљани а који су коришћени у раду [4], као ни све атрибуте који су ти слогови садржали, а који би применом неких новонасталих техника и другачијег софтвера од оног оригиналног аутора били употребљени за побољшање модела. Такође, додатни атрибути који се односе на годишњи/месечни приход самих контактираних особа или њихових породица би вероватно имао великог утицаја на исход кампање, тако да је то оно што би могло да буде у будућим кампањама прикупљано. Сем тога, претходно задовољство у коришћењу банкарских услуга и оцене на некој скали од 1-5 или 1-10 које би контактирани корисници давали је још један од параметара који би могао да буде кључан у креирању бољих и ефикаснијих предиктивних модела.

Још један битан фактор који је на резултате овог модела могао да утиче, а који кроз сам скуп атрибута није најбоље испраћен јесте велика економска криза која је погодила светску економију 2008. и 2009. године, а од које су се земље западне Европе потпуно опоравиле тек за годину-две, тако да људи можда у доба кризе и нису били спремни и довољно храбри на доношење нових битних финансијских одлука. Овај утицај финансијске кризе би могао да поквари реалност и применљивост модела у будућим ситуацијама, када је доба економске кризе превазиђено, тако да су потребни нови подаци новијег датума.

VII. Библиографија

[1] „Direct marketing,“ [На мрежи]. Available: http://en.wikipedia.org/wiki/Direct_marketing.

[2] C. L. J. H. N. Z. C. Ou, у *One Data mining for direct marketing*, Springer-Verlag Berlin, 2003, pp. 491-498.

[3] G. P. David Arnott, у *Eight key issues for the decision support systems discipline*, 2008, pp. 657-672.

[4] P. C. E. S. Moro, „A Data-Driven Approach to Predict the Success of Bank Telemarketing,“ *Decision Support Systems,* June 2014.

[5] H. Elsalamony, „Bank direct marketing analysis of data mining techniques,“ December 2013.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall и K. W. Philip, „SMOTE: Synthetic Minority Over-sampling Technique,“ *Journal of Artificial Intelligence Research 16,* pp. 321-357, 2002.

[7] R. B. P. W. A. L. Petrison, „Database marketing: Past, present, and future,“ *Journal of Direct Marketing,* pp. 11, 4, 109-125, 1997.