LOGISTIČKA REGRESIJA, KAO OSNOVA MAŠINSKOG UČENJA

Vlado Simeunović
Univerzitet Sinergija, Bijeljina, Republika Srpska,
v.simeunovic@spinter.net

Milan Milosavljević
Univerzitet Singidunum i Elektrotehnički fakultet,
Beograd, mmilan@etf.bg.ac.yu

Apstrakt: *Informaciono doba stvorilo je moderne kompjuterske i telekomunikacione tehnologije koje omogućuju prikupljanje i čuvanje ogromnih količina podataka. Sveprisutna elektronska i kompjuterska pomagala koriste se u praktično svim aspektima poslovnog i društvenog života. Osim što pojednostavljuju život, elektronska pomagala postaju generatori podataka, bilo da im je to osnovna namjena ili tek popratni efekt. Podaci nastaju i bilježe se u gotovo svakoj svakodnevnoj situaciji. Možda nismo ni svjesni koliki dio naših odluka, izbora i navika je zabilježen u različitim bazama podataka. Današnje tehnologije čine prikupljanje podataka jednostavnim, a njihovo skladištenje i čuvanje jeftinim. Stoga prikupljanje i čuvanje podataka prestaju biti problemi, a u focus dolazi njihova analiza i razumijevanje. Široka rasprostranjenost baza podataka stvorila je potrebu za snažnim analitičkim alatima koji sačuvane podatke mogu pretvoriti u korisne informacije. Postavlja se pitanje šta čini statističko-matematičku osnovu uspiješnosti mašinskog učenja.*
Ključne riječi: *logisitička regresija, data mining, logistička funkcija, baza podataka, linearna regresija, dihotomne promjenljive, mašinsko učenje*

Summary: *Information era has created modern computer and telecommunication technologies which allow collecting and recording huge amounts of data. All present electronically and computer gadgets are being used practically in all aspects of business and social life. Not only that they make the life less complicated, electronically gadgets are collectors of data, and it doest not matter it is the basic thing that it do, or just a side effect. Data are being made and being recorded in almost everyday situation. It might be that we are not aware how much our decisions, choices and various habits are being recorded in these various basis. Today technology make collecting and gathering of the technology simple, and their storage is very cheap. Therefore, collecting and storing the data are not problems any more, and the focus is now being moved to analysis and understanding. Wide spread usage of databases has created the need for strong analytical tools which saved data can use in various purposes. There is one question which can be asked in here: what makes the statistical and mathematical basis of the success in machine tech learning?*

UVOD

Regresija se koristi da ukaže na statistički metod koji omogućava predviđanje i ocjenjivanje jedne promjenljive na osnovu vrijednosti neke druge promjenljive ili više promjenljivih. Logistička regresija se može koristiti za predviđanje zavisne promjenljive na osnovu nezavisnih promjenljivih i određuje procenat varijanse u zavisnoj promenjivoj koja je definisana nezavisnim promenjivim; da bi rangirala relativnu važnost nezavisnih promjenljivih; da bi procijenila efekat interakcije; i da bi dokučila uticaj intervalne kontrolisane nezavisne promjenljive. Logistička regresija primjenjuje maksimalnu procjenu vjerovatnoće (MLE) nakon promjene zavisne promjenljive u logističku promjenjivu (prirodni

logaritam vjerovatnoća nezavisnih promjenljivih koje se pojavljuju ili ne). Na ovaj način logistička regresija procjenjuje vjerovatnoću da određeni događaj nastupi. Logstička regresija računa promjene u logaritmu vjerovatnoće zavisnih promjenljivih a ne promjene u zavisnim promenjivim kao što to radi OLS (ordinary least squares - kvadrat najmanje vrijednosti) regresija.

Logistička regresija ima mnogo analogija sa linearnom regresijem. Logaritamski koeficijent odgovara b koeficijentu u jednačini logističke regresije, standardizovani logistički koeficijent log odgovara beta težini, a pseudo R^2 statistika je dostupna da bi sumirala snagu odnosa. U našem radu smo testirali primjenu logisitčke regresije u kao alata u data miningu.

1. TEORIJSKE PRETPOSTAVKE KORIŠĆENJA LOGISTIČKE REGRESIJE

1.1. Pretpostavke za korišćenje logističke regresije

Primjena logističke regresije nije popularna među istraživačima jer zahtijeva ozbiljno i detaljno poznavanje mjernih skala, raspodjela i tumačenja rezultata. U istraživačkoj praksi ona omogućava istraživaču da prevaziđe mnoge pretpostavke koje ga ograničavaju linearnu regresiju i to:

1. Logistička regresija ne predpostavlja linearnu vezu između zavisnih i nezavisnih promenljivih. Može riješti nelinearne efekte čak i kada su eksponencijalni i polinominalni uslovi nisu eksplicitno dodati kao dodatne nezavisne vrijednosti.
2. Zavisna promenjiva bi trebala biti normalo distribuisana.
3. Zavisna promjenljiva mora biti homoskedastična za svaki nivo nezavisnih promjenljivih.
4. Obično uslovi distribuisane greške nisu pretpostavljeni.
5. Logistička regresija ne zahtijeva da nezavisne promjenljive budu interval.
6. Logistička regresija ne zahtijeva da nezavisne promjenljive budu neograničene.

Bez obzira na prethodno navedene prednosti, proces primjene logističke regresije zahtijeva ispunjenje određenih uslova i transformacija promjenljivih. Treba ispuniti sledeće uslove:

a) Tumačenje logističkih koeficijenata nije moguće bez smislenog kodiranja.

b) Uključenje svih relevantnih promjenljivih u regresioni model: Ukoliko su relevantne promjenljive ispuštene, uobičajena varijansa koju dijele sa uključenim promjenjivim može biti pogrešno pripisane ovim promjenjivim, ili uslovi greške mogu biti preuveličani.

c) Isključivanje svih irelevantnih promjenljivih ukoliko su uzročno i relevantne promjenljive uključene u model uobičjna varijansa koju dijele sa uključenim promjenjivim može biti pogrešno pripisane i relevantni promenjivim.

d) Predpostavlja se da su uslovi greške nezavisni (nezavisno prikupljanje).

e) Niska greška u promjenjivim. Idealno se pretpostavlja nisko mjerenje greške i slučajeva koji ne nedostaju.

f) Logistička regresija ne podrezumijeva vezu između zavisnih i nezavisnih promjenljivih kao što je to slučaj kod linearne regresije ali podrazumijeva linearnu vezu između logaritma zavisnih i nezavisnih promjenljivih. Box-Tidwell transformacija i ortogonalni polinominalni kontrast su načini testiranja lineranosti logaritma između nezavisnih promjenljivih.

g) Kao i linearna regresija, logistička regresija ne podrazumijeva efekte interakcije osim kad su uslovi interakcije kreirani kao dodatne promjenljive u anlizi.

h) Ne multikolinearnost: U izvijesnoj mjeri kada je nezavisna promjenljiva linearna

funkcije druge nezavisne promjenljive, problem multikolinearnosti će se desiti po logističkoj regresiji kao što se to dešava u linearnoj regresiji.

i) Veliki uzorci. Takođe, za razliku od OLS regresije logistička regresija maksimalno koristi likelihood procjene (MLE) radije nego OLS da bi izvela parametre. MLE se oslanja na asiptotičku normalnost velikog-uzorka što znači da pouzdanost procjena opada onda kada ima nekoliko slučajeva za svaku posmatranu kombinaciju nezavisne promjenljive.

j) Opravdanost okupljanju. Mjere dobrog uklapanja kao model hi kvadrata pretpostavlja da za ćelije formirane od kategoričkih nezavisnih promjenljivih sve ćelije frekvencije su >=1 i ne više od 20% ćelija su <5.

k) Očekivana disperzija. U logističkoj regresiji očekivana varijansa zavisne promjenljive se može uporediti sa posmatranom varijansom, i različitosti mogu biti smatrane na nivou ili iznad disperzije.

1.2. Matematički model logističke regresije

Jedna od matematičkih funkcija koja najbolje opisuje vezu između između promenljivih je funkcija logističkog odgovora:

$$\ln(\frac{P}{1-P}) = a + bX \qquad \frac{P}{1-P} = e^{a+bX} \qquad P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

ili

$$p_i = \frac{e^{(b0+b1x1+..+bvxv)}}{1+e^{(b0+b1x1+...bvxv)}}$$

gdje je *e* otprilike jednako 2.718, a *b* su težine logističke regresije (ili logaritamske težine) za nezavisnu varijablu. Ova metoda podrazumijeva dihotomne promjenljive date u binarnom obliku. Vrijednost Y=1 znači da se određeni događaj zaista dogodio, vrijednost Y=0 znači da je izostalo bilo kakvo događanje. X predstavlja nezavisnu promjenljivu ili mnoštvo atributa (X_1, X_2,....X_k).

Osnovna formula logistične regresije je:

$$P\{Y = 1/X\} = \frac{1}{1+e^{-x\beta}}$$

$$X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_k X_k$$

Funkcija $P(x) = \frac{1}{1+e^{-x}}$ je logistična funkcija. Ovu vrstu funkcije prvi je predstavio belgijski matematičar Vehulst. Na početku dvadesetog vijeka veći broj naučnika istraživalo je njene osobine i mogućnost upotrebe. Kao regresionu funkciju prvi je upotrijebio Koks 1958. godine. Kod logističke funkcije vrijednost je ograničena između 0 i 1. Iz osnovne postavke se vidi da se vrijednost 0 dobija u slučaju da je x = -∞ , a vrijednost 1 za x = ∞, za x = 0 vrijednost funkcije je 0,5.

Inverzni oblik logistične funkcije je logaritamska funkcija.

$$x = \log\frac{P}{1-P} = \log it(P)$$

Ako udružimo osnovnu formulu sa prethodnom dobićemo drugi oblik logistične regresije:

$$\text{logit}\,(P\{Y=1/X\}) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_k X_k$$

Diskretne atribute za više od dvije vrijednosti potrebno je pripremiti u formi više atributa. Pretpostavimo da je X_i diskretni atribut i ima q različitih vrijednosti od v_1 do v_q. U ovom slučaju jednačina se može napisati u sljedećem obliku:

$$\text{logit}(P\{Y=1/X\}) = X\beta = \beta_0 + ...$$

$$\beta_1 X_1 + \beta_2 X_2 + ..\lfloor \beta_{i2} I(X_i = v_2) + ..\beta_{iq} I(X_i = v_q) + ..\beta_k X_k \rfloor$$

Član $\beta i1I(X_i=v_1)$ je ispušten zbog rješavanja problema zavisnosti atributa.

Uspjeh logističke regresije može biti procijenjen provjerom u klasifikacionoj tabeli, prikazujući pravilne i nepravilne klasifikacije dihotomnih, uobičajenih ili politomnih zavisnih promjenljivih. Takođe, testovi mogućnosti uklapanja (prilagođavanja) kao što su model hi-kvadrat su dostupni kao indikatori modela primenljivosti kao što je Wald statistika da bi testirali važnost individualnih nezavisnih promjenjivih. Logistička regresija je tek u zadnje vrijeme postala popularnija među naučnicima, koji se bave društvenim naukama (ekonomija, bezbednost i sl.).

(Odds) se definiše kao odnosi mogućnosti i ne-mogućnosti određenog događaja.

$$odds = \frac{P}{1-P}$$

Binominalna (ili binarna) regresija je tip regresije koja se koristi kada je zavisna promenljiva dihotomna a nezavisna promjenljiva je bilo kog tipa.

Do sada smo uglavnom raspravljali o binominalnoj regresiji kako bi shvatili njene osnovne karakteristike. Naravno da se složeni problemi sa više od dvije promjenljive ne mogu riješiti na ovaj način te se primjenjuje multinominalna logisička regresija. Kada se višestruke grupe zavisnih promjenljivih mogu rangirati, tada je redna logistička regresija preformulisana u multinominalnu logističku regresiju.

Sledeći izraz predviđa mogućnost postojanja m kategorija u okviru zavisne varijable:

$$p(Yi = m) = \frac{1}{1 + \sum_{j=2}^{M} e^{(zij)}}$$

Gdje je Zij linearna funkcija koja se koristi da bi se predvideo m u okviru zavisne varijable (tj., b0 + b1x1+....bvxv).

2. METODOLOGIJA ISTRAŽIVANJA

2.1. Predmet i cilj istraživanja

Radi uočavanja mogućnosti primjene logističke regresije u okvirima mašinskog učenja, korisitli smo bazu podataka agencija koje se bave bezbednosnim problemima. Od mnogobrojnih oblika kriminalnog djelovanja od posebnog značaja za ugrožavanje sistema bezbednosti su: terorizam, trgovina ljudima, trgovina narkoticima, carinske prevare i krijumčarenja, krađe vozila, šverc oružjem za masovno uništenje, šverc lakog i malog naoružanja, finasijski i kompjuterski kriminal (krađa intelektualne svojine, falsifikovanje novčanica, pranje novca, kompjuterski kriminal, prevare kreditnim karticama). Bezbednosne agencije su strukturisane da svakodnevno prikuplja veliki broj podataka. Prikupljeni podaci se pohranjuju i čuvaju u različitim formama. U poslednje vrijeme, zbog ubrzane kompjuterizacije, većina dodataka se čuva u "bazama" podataka. Korišćenje podataka je i

dalje na relativno niskom nivou, uglavnom se operiše sa apsolutnim iznosima, frekvencijama, procentima, mjerama prosjeka i grafičkim prikazima koje prate deskriptivnu statistiku. Jedna od vrlo čestih kriminalnih radnji u poslednjih nekoliko godina bila je krađa motornih vozila. Svake godine nekoliko stotina motornih vozila bude ukradeno, pri čemu je stepen rasvijetljenosti vrlo nizak. Osnovni predmet istraživanja odnosi se na predlaganje primjene savremenih informatičkih sredstava i statističko-matematičkih procedura kao mehanizama koje olakšavaju predviđanje i donošenje odluke.

2.2 Hipoteze istraživanja

Osnovna hipoteza

Primjenom adekvatnih stastističkih metoda (klasičnih ili DM) mogu se uspješno predvidjeti tok, odvijanje, intenzitet, nastanak i nestanak određenih bezbednosnih pojava iz oblasti opšteg kriminaliteta.

Prva posebna hipoteza

Dosadašnji način obrade, interpretacije i primjene prikupljenih podataka iz oblasti opšteg kriminaliteta nije davao adekvatne rezultate u proaktivnom djelovanju ovlaštenih agencija.

Druga posebna hipoteza

Primjenom adekvatnih statističkih metoda mogu se uspješno klasifikovati, predvidjeti budući bezbjednosni događaji iz oblasti opšteg kriminaliteta.

2.3. Način istraživanja (Metode, tehnike i instrumenti) istraživanja

1.Metode: *statistički metod (logistička regrsesija), analiza sadržaja metod modelovanja, data mining.*

3. Obuhvat pojave (populacija i uzorak)

Populaciju čine svi događaji opšteg kriminaliteta (krađa motornih vozila) u 2004., 2005. i 2006. godini. U ovom istraživanju koristimo se potpunim obuhvatom pojave - populacija je ujedno i uzorak.

3. REZULTATI ISTRAŽIVANJA

3.1. Logistička regresija

Za dokazivanje mogućnosti primjene logističke regresije u predviđanje bezbednosnih događaja izradili smo i testirali odgovarajući model. Logističkom regresijom smo odredili procenat varijanse u zavisnoj promenjivoj koja je definisana nezavisnim promenjivim, zatim smo rangirali relativnu važnost nezavisnih promjenljivih; procijenili efekat interakcije i uticaj intervalne kontrolisane nezavisne promjenljive. Model obuhvata 707 slučajeva krađe motornih vozila u periodu od tri godine. Svi slučajevi su svrstani u dvije kategorije jedinstvene zavisne varijable koju smo nazvali "rasvijetljenost". Varijabla je dihotomnog karaktera gdje smo sa jedinicom obilježili sve "rasvijetljene" slučajeve, a sa dvojkom "nerasvijetljene". U ovom modelu imamo tipične diskretne vrijednosti i frekvenciju događaja. Nezavisne varijable su sledeće: marka vozila, dio grada gdje je vozilo nestalo, mesto parkiranja, registracija, osigurano, vrijeme parkiranja, vrijeme prijave nestanka. Svaka varijabla je predstavljena sa nekoliko kategorija.

Postoji nekoliko metoda procjene u logističkoj regresiji ali najčešći, možda i najmanje rizičan u smislu potvrđivanja hipoteze je METHOD=BSTEP(LR), za *Stepwise* analizu unazad. Metod se sastoji na mogućnosti testiranja log-likelihood-a sa datom promjenljivom ispuštenom iz jednačine.

U našem slučaju model smo testirali upoređujući početnu vrijednost logaritma koja iznosi 800,794 sa finalnim modelom koji iznosi 681,328. Uz 24 stepena slobode χ^2 iznosi 119,467 što je značajno na nivou ,000. Inicijalni hi kvadrat je devijansa za nulti model (model samo sa konstantom). To znači da je inicijalni hi kvadrat -2LL za model koji prihvata nultu hipotezu da su svi β koeficijenti 0. Ovo podrazumjeva da ni jedna od nezavisnih promenljivih nije linearno povezana sa logaritamskim vjerovatnoćama zavisnih promenljivih.

χ^2 = -2LLinicijalni - (-2LLfinalni) = 800,794 - 681,328 = 119,467

Tabela br.1: Logaritam izračunatih vjerovatnoća

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	800,794	119,467	24	,000
Final	681,328			

Dobijeni rezultati govore da je model smislen i da se ne može prihvatiti nulta hipoteza o nepostojanju povezanosti između nezavisnih i zavisne varijable. Model Hi kvadrat procjenjuje sveobuhvatni logistički model, ali nam ne govori da li su određene nezavisne promjenjive značajnije od drugih. U sledećem koraku smo pokazali stepen povezanosti pojedinih varijabli i strukturu konačnog modela. U tabeli br. 2 prikazan je nivo obuhvaćene varijanse modelom koj iznosi 70,7 %. Nulti model se može predstaviti odnosom 207/500 = 70,7

Tabel br.2: Klasifikaciona tabela

Posmatrane		Predviđene		
		RASVIJETLJENO		
		ne	da	
Korak 0 RASVIJETLJENOST				
	ne	500	0	70,7
	da	207	0	29,3
Ukupni postotak				100,0

U našoj logističkoj regresiji BSTEP(LR) na početku sve promjenjljive ušle u jednačinu, a zatim je model testiran u šest koraka. Kroz postupak od šest koraka hi – kvadrat test se postepeno smanjivao, što je i imperativ modela, tako da je da smo od početne vrijednosti 66,705 dobili umanjenu vrijednost koja iznosi 62,510. Prva varijabla koja je eliminisana iz modela je "osiguranje" -,067.U sledećem koraku eleminisana je varijabla "mjesto parkiranja" sa koeficijentom koji se u drugom koraku povećao na vrijednost ,054. U trećem koraku eleminisana je varijabla "vrsta zaštite" sa konačnom vrijednošću koeficijenta ,073. U četvrtom koraku učešće varijable "dio grada" nije značajno učestvovalo u poboljšanju vjerovatnoće ukupnog modela (,158). Zadnja varijabla koja nije učestovala u završnom modelu je "raspisana potraga" sa koeficijentom -,258. U konačnom modelu učestvuju varijable marka vozila sa koeficijentom (,132), zatim varijabla "registracija" sa koeficijentom (-1,035), varijabla ‘’vrijeme parkiranja’’ (,230) i varijabla ‘’vrijeme prijavljivanja’’ (-,342). U finalnom modelu Cox & Snell iznose 0,143, a korekcijom Nagelkerke indeksom dobija se vrijednost 0,204 što se može smatrati zadovoljavajućim isahodom.

Tabela br.3 Sumarni model

Step	- 2 Log likelihood	Cox&Snell R Square	Negelkerke R Square
1	820,640	,047	,068
.			
6	745,742	,143	,204

U konačnom modelu od ukupno 500 slučajeva koji su trebali imati vrijednost 0, samo 31 slučaj dobio je vrijednost 1, a od 200 slučajeva koji su trebali uzeti vrijednost 1, 43 slučaja je uzelo vrijednost 0. Procenat pravilnog klasifikovanja je 72,4%. Sveobuhvatni model promjenljivih koje ulaze u jedančinu logističke regresije posmatran je preko vrijednosti B, S.E., Wald, df, Sig., Exp(B), C.I.. Koeficijent B za ukupan model iznosi:

$$\ln(\frac{p_i}{1-p_i}) = b_0 + b_1 x_1 + ... + b_v x_v$$ =,397+,132*374-1,035*394+,230*270 +,342*289 = -282,898

Da ne bi ostali na površini kvantitativnih podataka izvršili smo detaljniju statističko-matematičku analizu svakog podatka. Pomoću naredne jednačine izračunali smo vjerovatnoću svake promjenljive koja značajno utiče na stepen otkrivanja krivičnog djela (zavisna promjenljiva).

$$P(x) = \frac{e^{b0+b1*x}}{1+e^{b0+b1*x}}$$

U petom koraku dobili smo vrijednost konstante b_0 (b_0 = 0.397). Konstanta ulazi u eksponencijalnu funkciju kao prvi član. Izračunali smo vjerovatnoću svake varijable koja je ušla u jednačinu. Najveća vjerovatnoću P(x)=0.861 ima prediktor – vrijeme prijavljivanja, čiji je b koeficijent 0.230, zatim slijedi prediktor marka vozila (P(x)=0.8345; b1=0.132); vrsta registracije je po težini na trećem mjestu u jednačini (P(x)=0.7056 b1=-1.035) a vrijeme parkiranja na četvrtom mjestu (P(x)=0.5925 b1=-0.342). Bez obzira što smo identifikovali varijable koje sa visokom vjerovatnoćom predviđaju ponašanje zavisne promjenljive, nužno je napraviti još jedan korak u pojedinačnom identifikovanju učešća pojedinih kategorija u prediktorskim varijablama. Najveću vjerovatnoću od svih kategorija u prediktorskim varijablama ima ‘’marka vozila 5’’ (P(mv(5)) = 0.8552 b1=0.662). Greška odstupanja za ovu varijablu iznosi S.E. = 0,288; Wald $Wv = (bv / Sbv)^2$=5,286; Exp(B)= 1,939 sa koeficijentom povjerenja 95,0% C.I = $e^{\beta_0 \pm 1,96SE\beta}$ =(1,103 - 3,410). Iz navedenih rezultata jasno se vidi da je vjerovatnoće krađe vozila VW najveća, a da je dva puta veća vjerovatnoća da ovo auto ne bude nađeno nego da bude nađeno. ‘’Marka vozila 4" ima visoku vjerovatnoću koja se kreće u granicama (P(mv(4)) = 0.8529, b1=1,188). Radi se o vozilima marke mercedes čija vrijednost na tržištu je velika. Greška odstupanja za ovu varijablu iznosi S.E. = 0,469 sa stepen značajnosti Sig. = 0,011. Vrijednost Exp(B)= 3,280, 95,0% C.I = 1,307 - 8,232, ukazuje da je mogućnost pronalaska otuđenih vozila navedene marke 3 puta manja od mogućnosti da se vozilo pronađe i vrati vlasniku. Veoma indikativan rezultat se odnosi na varijablu "registracija 1". Iznos vjerovatnoće je veoma visok [P(reg(1)) = 0.8288, b1=-0.952]. Koeficijent b1 je negativan u odnosu na nerasvijetljene slučajeve, što znači ukoliko se povećava broj ukradenih vozila domaće registracije utoliko se povećava broj otkrivanja. Očegledno da ovaj podatak ima smisla za poznavaoce prilika u sektoru bezbednosti. Za stvaranje ukupne slike važan je podatak da je broj ukradenih vozla strane registracije nesrazmjerno veliki u odnosu na njihov broj (394 vozila sa domaćom registracijom, naspram 313 vozila sa stranom registracijom strana). Nesrazmjernost je uočljiva i u ukupnom broju rasvijeteljenih slučajeva. Od 394 slučajeva krađe vozila sa domaćom registracijom, 150 slučajeva je rasvijetljeno ili 38 % u

odnosu na 57 slučajeva krađe vozila sa stranom registracijom od 313 slučajeva ili oko 18%. Očigledno je teža potraga za vozilima strane registracije jer bezbednosne agencije imaju manje podataka o njima, ali je i sistem osiguranja takav da vlasnici vozila vrlo često ne žure sa prijavom krađe jer je mogućnost naplate šetete od osiguravajućih društava izvijesan. ‘’Vrijeme parkiranja 1’’ je veoma značajna varijabla u jednačini čija je vjerovatnoća veoma visoka i iznosi ((P(vpar(1)) = 0.9078 b1=0.785). Radi se o periodu od 02.00 do 06.00, što samo po sebi govori. Greška odstupanja za ovu varijablu iznosi S.E. = 0, 417 sa stepen značajnosti Sig. = 0,002. Vrijednost Exp(B)= 2,191, ukazuje da je mogućnost pronalaska otuđenih vozila otuđenih u ovom vremenskom periodu dva puta manja od mogućnosti da se vozilo pronađe i vrati vlasniku sa stepenom povjerenja 95,0% C.I =,780-3,994.

Rezultati krostab analize pokazuju da nije najveći broj vozila nestalo u ovom vremenskom periodu (138 od ukupno 707), ali je broj riješenih slučajeva 22. Ovaj podatak ima jasnu informacionu vrijednost jer se radi o noćnom vremenu. Jedna od varijabli koja ulazi u jednačinu je postupak koji se odnosi na rad bezbednosnih agencija, a u vezi je sa vremenom prijavljivanja nestanka vozila. Od ove aktivnosti najčešće zavisi brzina rješavanja problema. Jedna kategorija navedene varijable statistički značajna na nivou ,000. Radi se o varijabli ‘’vrijeme prijavljivanja 2’’ čija je vjerovatnoća P(vpri(2)) = 0.6386. Koeficijent logističke regresije iznosi (b1=-0.903). Navedeni koeficijent je negativan što znači da je najveća vjerovatnoća rješavanja bezbednosnog događaja ukoliko se izvrši prijava u vremenu od 06,00 do 10,00. Sasvim je logično da je ovo vrijeme koje pruža najveću mogućnost angažovanja svih kapaciteta bezbednosne agencije u aktivnostima pokretanja potrage jer su službenici na svojim radnim mjestima. Greška odstupanja za ovu varijablu iznosi S.E. = 0, 237, vrijednost Exp(B)= ,406 ukazuje da se sa stepenom povjerenja 95,0% C.I =,780-3,994 može tvrditi da će pronalazak otuđenih vozila za 40% biti veći ako se nestanak prijavi u periodu od 06.00 do 10.00. Frekvencija rasvijetljenih slučajeva iz ove kategorije iznosi 107 što je preko pedeset posto svih riješenih slučajeva. Wald test iznosi 14,521 te je odbačena nulta hipoteza po kojoj se očekivao rezultata 0.

4. **ZAKLJUČNA RAZMATRANJA**

U radu smo u osnovi pokušali da primijenimo neke od statističkih postupaka za predviđanje parametara značajnih za odlučivanje u situacijama gdje ti parametri nisu sređeni, a dati su u opisnom (atributivnom) obliku. Osnovni cilj istraživanja je pokazati da se na osnovu relativno velikog broja podataka, koji se godinama skupljaju u bezbednosnim agencijama uspostavi sistem i formira baza podataka sa dovoljnim brojem kvalitetno definisanih atributivnih varijabli koje bi primjenom neke od klasičnih statističkih procedura ili data mininga mogle poslužiti za klasifiklaciju ili predviđanje određenih događaja. Na ovaj način bi se obogatio postojeći teorijski fond znanja u primjeni statistike u bezbednosnim naukama, ali i riješili određeni praktični problemi proaktivnog djelovanja. Osim navedenog cilj nam je bio da uočimo mjesto logisitičke regresije u procesima data m ininga. Kao najprimjerenija statistička tehnika analize atributivnih varijabli korišćena je logistička regresija. Procenat obuhvaćene varijanse koja se može predvidjeti logističkom regresijom iznosi 70,7%. Koristeći METHOD=BSTEP(LR), za *Stepwise* logističku analizu unazad u petoj iteraciji izdvojile su se sledeće varijabel: marka vozila, vrijeme parkiranja i vrijeme prijavljivanja nestanka vozila. Svaka od tri navedene varijable su statistički značajne na nivou 0,01 i većem i to: marka vozila (0,01), vrijeme parkiranja (0,002) i vrijeme prijavljivanja (0,000). Radi se o varijablama koje su ušle po strukturi u jednačinu logističke regresije i za koje se može smatrati da predstavljaju prediktore u procesu odlučivanja. Korisnicima ovog i sličnih radova navedeni podaci mogu poslužiti da svoje aktivnosti usmjere u pravcu formiranja modela odlučivanja koji bi obuhvatio navedene varijable. Iz analiza se može

zaključiti da su vozila poznatih svjetskih marki najčešće na meti kriminalnih grupa, a posebno VW, audi i mercedes. Varijabla "vrijeme parkiranja" govori o tome da se krađe događaju obično u kasnim noćnim satima, a da proces pronalaženja zavisi od brzine prijavljivanja slučaja. Logistička regresija se pokazala kao moćan alat za pronalaženje zakonitosti u skupu nominalnih varijabli. Čini se da ima poseban praktični značaj, ali je nužno formirati baze podataka sa jasno definisanim atributima, sa što više kategorija, kako bi diskriminativna vrijednost rezultata bila veća. Jedan od problema koji se pojavljuju u primjeni logističke regresije predstavlja tumačenje pojedinih rezultata koji su dati u razlčitim matematičko-statitičkim formama (logaritam, vjerovatnoća, koeficijenti koji nisu u dovoljnoj mjeri razumljivi npr.Cox i Snell i sl.). Primjena logisitičke regresije daje najbolje rezultate, ukoliko ona predstdavlja alat u data miningu.

LITERATURA

1. Additional Protocol to the European Convention on Mutual Assistance in Criminal Matters, Strasbourg, 17.III.1978.
2. Allison, Paul D. (1999). Comparing logit and probit coefficients across groups. Sociological Methods and Research, 28(2): 186-208.
3. Cander L.R., Stankovic S., Milosavljevic M. (1997). Dynamic ionospheric prediction by neural networks, AI Applications in Solar-Terrestrial Physics.
4. Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psych. Bull., 70, 213-220.
5. Cox, D.R. and Snell, E. J. (1989). Analysis of binary data (2nd edition). London: Chapman & Hall.
6. Criminal codes and Criminal procedure codes from SEE countries.
7. Fajgelj, S. (1995). Faktorska analiza nominalnih varijabli, Magistarska teza, Filozofski fakultet, Beograd.
8. Green, A.M. (1997). Kappa statistics for multiple raters using categorical classifications. Proceedings of the 22nd annual SAS User Group International conference, 1110-1115.
9. Hosmer D.W., Lemeshow, S. (2000). Applied logistic regression. Wiley & Sons, New York,
10. Ian, H. Witten and Eibe Frank, (2002)."WEKA Tutorial", University of Waikato.
11. Ian H. Witten and Eibe Frank, (2000). “Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations”, Morgan Kaufman.
12. Kleinbaum, D. G. (1994). Logistic regression: A self-learning text. New York: Springer-Verlag. What it says.
13. Kononenko, I., (1992). Naive Bayesian classifier and continuous attributes. Informatica, 16(1):1–8.
14. Krstić, Ostoja: Kriminalistička prevencija i prognostika, Visoka škola unutrašnjih poslova, Banja Luka, 2005.
15. Menard, S. (2002). Applied logistic regression analysis, 2nd Edition. Thousand Oaks, CA: Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 106. First ed., 1995.
16. Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. Biometrika, Vol. 78, No. 3: 691-692. Covers the two measures of R-square for logistic regression which are found in SPSS output.
17. Preradović, Lj., Simeunović, V. (2007). Informatika-Metodologija-Statistika, Visoka škola unutrašnjih poslova, Banja Luka.

18. Second Additional Protocol to the European Convention on Mutual Assistance in Criminal Matters, Strasbourg, 8.XI.2001
19. Security and combating international organized crime and terrorism, W. Bruggeman. December 2002,

Internet:

http://personal.ecu.edu/whiteheadj/data/logit/
http://www2.chass.ncsu.edu/garson/pa765/logistic.htm