Primjena strojnog učenja za određivanje rizika davanja kreditne kartice

Slunjski, Tomislav

Master's thesis / Diplomski rad

2019

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Organization and Informatics / Sveučilište u Zagrebu, Fakultet organizacije i informatike**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:211:281832

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Tomislav Slunjski

Primjena strojnog učenja za određivanje rizika davanja kreditne kartice

DIPLOMSKI RAD

Varaždin, 2019.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Tomislav Slunjski

Matični broj: 43564/14–I

Studij: Informacijsko i programsko inženjerstvo

Primjena strojnog učenja za određivanje rizika davanja kreditne kartice

DIPLOMSKI RAD

Mentor/Mentorica:

Doc. dr. sc. Dijana Oreški

Varaždin, rujan 2019.

Tomislav Slunjski

Izjava o izvornosti

Izjavljujem da je moj diplomski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor potvrdio prihvaćanjem odredbi u sustavu FOI-radovi

_____

Sažetak

U ovom radu koriste se pristupi strojnog učenja za rješavanje problema identifikacije potencijalno loših klijenata u bankarskom sektoru prilikom dodjele kreditnih kartica. Jednom od tehnika strojnog učenja identificiraju se pravila dodjele kartice koja se implementiraju u sustav temeljen na znanju. Sustav se koristi kao potporu odlučivanju u domeni bankarstva.

Kako su banke u svome poslovanju izložene kreditnom riziku i cilj im je reducirati kreditni rizik, ovaj sustav bi im uvelike pomogao u ostvarivanju toga cilja. Jer u suprotnom dolazi do krive procjene koja za sobom povlači gubitke koji nikako nisu u interesu banke i potrebno ih je svesti na minimum.

Već postoje razvijeni modeli kreditnog skoringa, a u ovom radu bi se razvio sličan model na temelju stabla odlučivanja. Za kvalitetan model bitno je dobro odrediti karakteristike ključne za predikciju kreditne sposobnosti klijenta banke.

Ključne riječi: stablo odlučivanja; model; procjena rizika; strojno učenje; ekspertni sustav; banka; kreditne kartice;

Sadržaj

1. Uvod 1

2. Kreditni rizici 2

3. Metode i tehnike rada 5

3.1. Stablo odlučivanja 6

3.2. Simulacija podataka 9

3.3. Softverski alat za primjenu metode stabla odlučivanja 11

4. Analiza prethodnih istraživanja 13

4.1. Primjena stabla odlučivanja u kreditnom skoringu 13

4.2. Procjena kreditnog rizika pomoću statističkog i strojnog učenja: Osnovne metodologije i aplikacije za modeliranje rizika 18

4.3. Potrošački modeli kreditnog rizika putem algoritama strojnog učenja 21

5. Metoda stabla odlučivanja za određivanje rizika davanja kreditne kartice 23

5.1. Opis problema 23

5.2. Rješavanje problema primjenom metode stabla odlučivanja 24

5.2.1. Skup podataka 24

5.2.2. Izrada modela 27

5.2.3. Evaluacija modela 30

5.2.3.1. Generalni rezultati evaluacije 32

5.2.3.2. ROC krivulja 36

5.2.3.3. Krivulja preciznosti odziva 37

5.2.3.4. Krivulja informacijskog dobitka 38

5.2.3.5. Kolmogorov-Smirnov statistika 39

5.2.3.6. Krivulja podizanja 40

5.2.4. Rad i primjena modela 41

6. Zaključak 46

Popis slika 50

Popis tablica 52

1. Uvod

Kvesić [1] opisuje kreditni skoring kao sustav dodjeljivanja bodova zajmotražitelju pri čemu se primjenjuje statistička analiza ključnih karakteristika zajmotražitelja s ciljem numeričke kvantifikacije rizika koja pokazuje vjerojatnost da će klijent doći u status neispunjavanja ugovorene obveze. Navodi da kreditni skoring sustav dodjeljuje jednu kvantitativnu mjeru, takozvani skor, potencijalnom klijentu predstavljajući njegovo buduće ponašanje u otplati dodijeljenoga kredita. Ukupan broj bodova (skor) dobiven sumom bodova po određenim karakteristikama uspoređuje se s najniže prihvatljivim brojem bodova definiranim kreditnom politikom banke na temelju čega se kredit odobrava ili ne odobrava. U njezinom radu navode se 2 vrste skoring modela, a to su aplikativni (koji se odnosi na nove klijente) i bihevioralni (koji se odnosi na postojeće klijente).

U ovom radu posvetit ćemo se izradi aplikativnog modela stabla odlučivanja koji bi se koristio kao zaseban ili dio sustava za potporu odlučivanju, točnije za procjenu kreditne sposobnosti novih klijenata, tj. da li klijent zadovoljava uvjete za dodjelu kreditne kartice. U tu svrhu koristit će se varijable demografskog i financijskog karaktera. Prilikom istraživanja prikupio sam nekoliko formi, različitih kartičnih kuća koje posluju na području Republike Hrvatske, koje se ispunjavaju prilikom zahtjeva za izdavanje kreditne kartice. Na temelju njih odredio sam varijable koje mogu biti od značaja za odluku o dodjeljivanju kreditne kartice te simulirao podatke koje sam koristio u svrhu izrade modela stabla odlučivanja. Uz izradu praktičnog dijela rada, cilj je i objasniti teorijski dio vezan uz temu.

Istraživanjem ove teme naišao sam na nekoliko prethodnih istraživanja, o kojima ću pisati detaljnije u radu, koja su istaknula metodu stabla odlučivanja kao najefikasniju. Stoga sam se odlučio koristiti tu metodu za izradu modela na kojem bi se temeljio ekspertni sustav.

2. Kreditni rizici

Kroz ovo poglavlje predstavit ću teoretski okvir za kreditne rizike. Jer je jedan od ciljeva ovog rada smanjenje i procjena kreditnog rizika.

Kada financijska institucija odobrava kredit zajmotražitelju, ona obavlja svoju fundamentalnu funkciju: prihvaćanje kreditnog rizika. Uspjeh banke leži u njenoj sposobnosti da predvidi i kvantificira ukupan rizik. Bankovni menadžment se suočava sa šest temeljnih rizika: kreditni rizik, rizik likvidnosti, tržišni rizik, operacijski rizik, regulatorni rizik i rizik ljudskog faktora [2, p. 62]. Fokusirat ću se na kreditni rizik jer je on u domeni ovoga rada.

Kreditni rizik je posljedica ugovorene i/ili moguće financijske transakcije između davatelja i uzimatelja sredstava odnosno varijacija mogućih povrata koji bi se mogli zaraditi na financijskoj transakciji zbog zakašnjelog ili nepotpunog plaćanja glavnice i/ili kamate [2, p. 63].

Svaki puta kada osoba ili poduzeće uzimaju neki proizvod ili uslugu bez trenutačnog plaćanja tog proizvoda ili usluge smatramo da stvaraju obvezu naknadnog plaćanja, tj. ostvaruju kredit. Naravno, postoji vjerojatnost da neće vraćati kredit, konkretno da neće podmirivati obveze po kreditu prema ugovoru o kreditu, no strana koja izdaje robu ili uslugu mora biti toga svjesna te time prihvaća kreditni rizik.

Krediti su danas dostupni u raznim oblicima pa tako i u obliku kreditne kartice koje pretežno nude banke kao opciju dinamičnijih i manjih kredita sa kraćim rokovima otplate pretežno na mjesečnoj bazi. Takvi oblici kreditiranja su dosta popularni pa čine i jedan dio poslovanja banaka, u Hrvatskoj čak značajan dio. Banka se mora zaštititi od rizika jer svako zakašnjenje ili izostanak povrata ugovorene glavnice i kamata smanjuje dobitak banke, a ujedno njenu stvarnu vrijednost. Razni faktori utječu na kašnjenje u otplati kredita.

U svrhu zaštite od kreditnog rizika, banke uvode kreditne procese koji se sastoje od kreditne analize, odluke o odobravanju kredita i nadgledanju kredita. Takav proces omogućuje menadžmentu i kreditnim referentima procjenu rizika. U kreditnoj analizi mjeri se sposobnost i potreba klijenta kako bi se utvrdilo da li je odabrani način financiranja pogodan. Rezultat kreditne analize je vjerojatnost hoće li ili neće klijent otplaćivati prispjele obveze po kreditu. Kreditna analiza započinje podnošenjem kreditnog zahtjeva uz koji idu i financijski i kreditni izvještaji tražitelja kredita. Zatim se provodi analiza dospjelih financijskih izvještaja i tijeka gotovine i procjena kolaterala. Kada se završe sve procjene i analize daje se prijedlog za odobravanje/odbijanje kreditnog zahtjeva. Izdvojit ću ključne faktore rizika prema Šarliji [2], a to su:

- Karakter: Klijent treba imati jasno definiranu namjenu kredita i ozbiljnu namjeru da kredit vraća prema ugovorom utvrđenom načinu.
- Kapital: Klijent treba biti sposoban voditi poslovanje koje će proizvesti takav novčani tijek koji će biti dovoljan za podmirenje svih obveza uključujući i otplatu kredita.
- Kapacitet: Klijent treba imati pravnu sposobnost i osposobljeni menadžment da vodi cjelokupno poslovanje.
- Kondicije: Podrazumijeva analiziranje okruženja klijenta koje utječe na definiranje uvjeta i pretpostavki za njegovo poslovanje.
- Kolateral: Kolateral je sekundarni izvor naplate potraživanja, a nastupa kada se pokaže da klijent nije u stanju ispuniti svoje obveze po kreditu.
- Kontrola: Podrazumijeva pregled zakonske regulative kako bi se otkrilo na koji način ona može utjecati na financijski položaj klijenta.

U inicijalnoj kvantitativnoj analizi provodi se:

- analiziranje kreditne povijesti
- ocjenjivanje unutarnjih i vanjskih faktora poduzeća
- analiziranje financijskih izvještaja
- analiziranje tijeka gotovine
- pravljenje izvještaja o kreditu

Kako navodi Šarlija [2], prijedlog za odobravanje/odbijanje kreditnog zahtjeva moguće je dobiti na dva načina. Jedan je subjektivna ocjena koja se dobije kada analitičar prilikom analize ključnih faktora rizika donosi svoju odluku isključivo na temelju znanja, iskustva i sljedeći pri tome propisane procedure i pravila. Drugi način je korištenje modela koji je razvijen upotrebom statističkih i ostalih metoda, a nazivamo ga kreditni skoring. To je sistem dodjeljivanja bodova zajmotražitelju čiji zbroj predstavlja numeričku vrijednost na temelju koje se može donijeti jasna odluka da li se kredit može odobriti ili ne. Nakon ispitivanja kreditne sposobnosti klijent se raspoređuje u jednu od rizičnih kategorija koje označavaju odobravanje ili odbijanje kreditnog zahtjeva. Odobravanje kredita kao posljedicu ima izradu dokumentacije i isplatu novčanih sredstava.

Završni dio kreditnog procesa je nadgledanje kredita. Sama riječ govori da se ovdje radi o praćenju kredita i otplate u skladu s potpisanim ugovorom. Moguće je i poduzimanje korektivnih aktivnosti kao što su promjena kreditnih uvjeta, dodatno osiguranje i slično. Glavni cilj je smanjenje kreditnog rizika i rješavanje problematičnih kredita. Proces nadgledanja je zanimljiv jer se i u njemu mogu koristiti skoring modeli, kako sam naveo postoje aplikativni, koji se odnose na nove klijente, i bihevioralni skoring modeli, koji se odnose na postojeće klijente i njihove račune. Kako navodi Šarlija [2] u aplikativnim skoring modelima za građane, karakteristike koje se koriste su sljedeće: vrijeme provedeno na postojećoj adresi stanovanja, stambeni status, posjedovanje telefona, godišnji prihod, posjedovanje kreditnih kartica, tip bankovnog računa, dob, zanimanje, namjena kredita, bračni status, trajanje računa u banci, stabilnost zaposlenja, podaci kreditnog biroa, mjesečne obveze, mjesečni prihodi, broj osoba u domaćinstvu itd. Većina navedenih varijabli pokrivena je i u simulaciji skupa podataka korištenoj u ovome radu. Dok se u bihevioralnima koriste i aplikativni podaci te podaci ponašanja vezani za postojeći račun klijenta. Odabir samih karakteristika koje će se koristiti u modelu najviše zavisi o tipu, količini i namjeni kredita.

3. Metode i tehnike rada

Kako bi mogao krenuti u izradu ovog rada potrebno je bilo istražiti i proučiti literaturu i znanstvene radove koji se bave sličnom tematikom. Prilikom istraživanja, naišao sam na tri znanstvena rada o kojima ću detaljnije pisati u nastavku ovog poglavlja. Ti radovi su me usmjerili na proučavanje određenog teoretskog dijela vezanog za temu. Konkretno na metodu stabla odlučivanja.

Uz to sam kroz sugestiju mentorice krenuo u proučavanje alata BigML. Radi se o alatu dostupnom putem web sučelja, koji je dosta intuitivan i besplatan u edukacijske svrhe što ga čini idealnim kandidatom za izradu potrebnog modela.

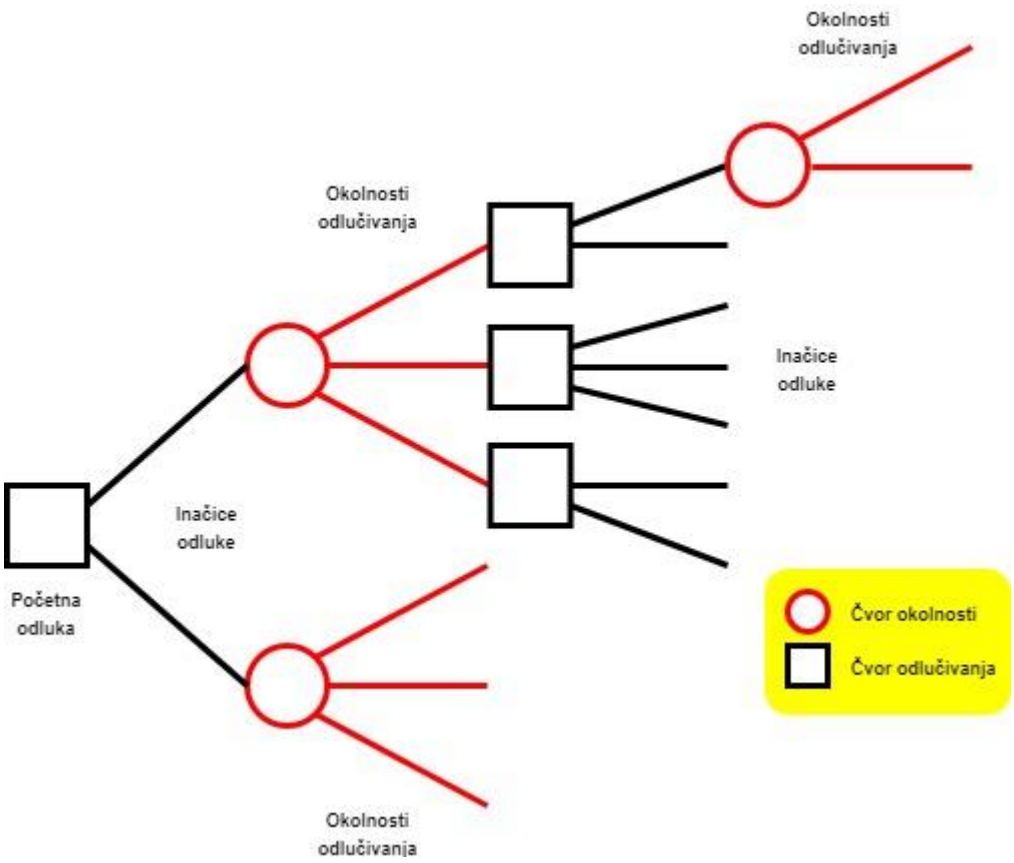
Kako je za izradu modela potreban određen skup podataka, pokušao sam na relevantnim internetskim stranicama pronaći potreban skup, ali rezultati mi nisu bili zadovoljavajući pa sam se odlučio na simulaciju podataka kako bi dobio skup podataka koji mi odgovara i koji je zadovoljavajući za potrebe ove teme.

Sve navedeno ću detaljnije opisati u nastavku.

3.1. Stablo odlučivanja

Kako sam naveo na početku poglavlja, proučavanjem znanstvenih radova na sličnu temu došao sam do zaključka da će mi za potrebe diplomskog rada najviše koristiti metoda stabla odlučivanja. U radovima se navodi kako se ta metoda pokazala najpreciznijom prilikom donošenja odluka koje su slične odluci kojom se bavi ovaj rad. To me nagnalo na korištenje ove metode stoga ću u nastavku predstaviti tu metodu kroz teorijski dio.

Sikavica i sur. [3] navode kako se stablo odluke u odlučivanju primjenjuje kao grafički model za vizualizaciju procesa odlučivanja kad se rješavanje problema odlučivanja svodi na donošenje više sukcesivnih odluka, a uz takav prikaz problema odlučivanja veže se i postupak računanja očekivanih vrijednosti inačica odluke u uvjetima rizika.

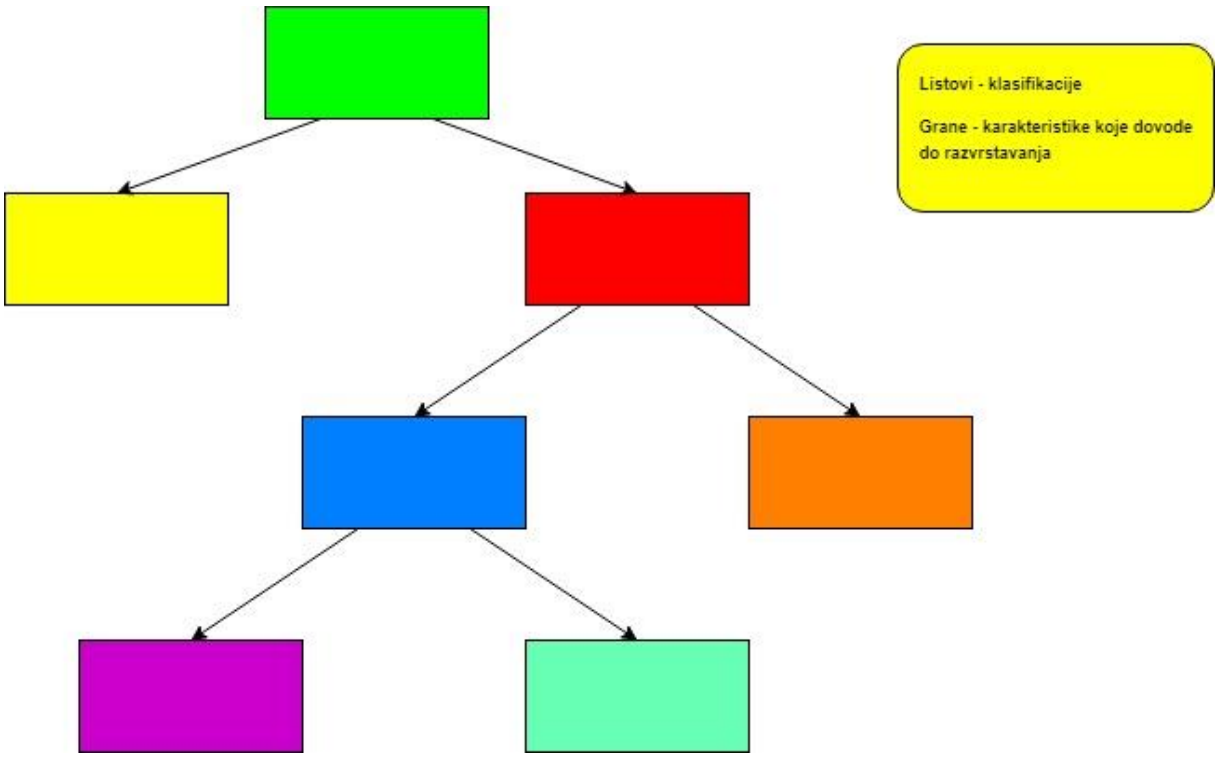


Slika 1. Grafički prikaz stabla odluke [4]

Stablo odlučivanja pogodno je za situacije u kojima se donošenje odluke temelji na vremenski slijedno povezanom nizu manjih odluka. Uglavnom se primjenjuje u poslovnom svijetu kada je potrebno donijeti odluku u rizičnim situacijama, a pri donošenju odluka stablom odlučivanja koristimo se kriterijem očekivane vrijednosti. Njegov grafički prikaz jasno prikazuje sve mogućnosti i definira problem odlučivanja. Započinje točkom odlučivanja nakon koje slijede alternativne odluke koje možemo donijeti. Svaka alternativna odluka prikazana je svojom vlastitom granom koja izlazi iz točke odlučivanja.

Sikavica i sur. [3] navode da se na kraju svake grane nalaze ishodi rizičnih slučajeva na temelju kojih se računa očekivana vrijednost čvora slučaja. Kod donošenja odluke bira se grana koja vodi prema čvoru s najvećom očekivanom vrijednošću jer se potpuna (savršena) informacija ne može dobiti pošto je to podatak koji se odnosi na budućnost.

Zekić – Sušac [5] navodi da u strojnom učenju, stabla odlučivanja su prediktivni modeli koji na temelju podataka izvode njihove veze u cilju dobivanja izlaznih vrijednosti. Kao takvi modeli koriste se u rudarenju podataka tražeći skrivene veze među podacima. Takva se stabla temelje na podacima, a ne na odluci eksperta i zovu se klasifikacijska i/ili regresijska stabla. Metodom klasifikacijskih i regresijskih stabala dobiva se grafički prikaz modela utjecaja ulaznih varijabli na izlaznu, pri čemu izlazna varijabla mora biti izrađena u obliku klasa ili kategorija.



Slika 2. Klasifikacijsko stablo odlučivanja [5]

Zekić – Sušac [5] opisuje način funkcioniranja stabla na način da:

- Svaki čvor u grafičkom stablu predstavlja jednu ulaznu varijablu
- Svaki list u stablu predstavlja vrijednost ciljne varijable ako su dane vrijednosti ulaznih varijabli predstavljene putom od korijena stabla do tog lista
- Stablo se dobiva „učenjem na podacima, na način da se vrši grananje izvornog skupa podataka u podskupove na temelju testiranja vrijednosti varijabli
- Proces se ponavlja na svakom izvedenom podskupu na rekurzivni način. Rekurzija je završena kada podskup određenog čvora ima sve iste vrijednosti izlazne varijable ili kada daljnje grananje više ne doprinosi poboljšanju rezultata

Kvesić [1], koristeći stabla odlučivanja u domeni strojnog učenja, daje užu definiciju više vezanu za temu ovog rada, a to je da stabla odlučivanja predstavljaju ne parametarsku tehniku klasifikacije klijenata u skupine, gdje je svaka skupina homogena u odnosu na rizik neispunjenja obveza i maksimalno diferencirana od rizika neispunjenja obveza ostalih skupina. Stabla odlučivanja vrlo su moćne i popularne tehnike za klasifikacijske i predikcijske probleme.

Prema Kvesić [1], stablo odlučivanja predstavljeno je klasifikacijskim algoritmom u obliku razgranatog stabla s čvorovima. Postoje dvije vrste čvorova:

- krajnji čvor (njime završava određena grana stabla)
- čvor odluke (definira određeni kriterij u obliku vrijednosti atributa iz kojeg izlaze grane koje zadovoljavaju određene vrijednosti tog atributa)

3.2. Simulacija podataka

Za simulaciju podataka odlučio sam se iz razloga što na relevantnim internetskim izvorima nisam pronašao skup podataka koji bi odgovarao potrebama ovoga rada. Kako bih dobio varijable na temelju kojih bi simulirao podatke, u fazi istraživanja za rad, prikupio sam pristupne forme, koje se ispunjavaju prilikom predaje zahtjeva za dodjelu kreditne kartice, nekoliko kartičnih kuća koje posluju na prostoru Republike Hrvatske. Na temelju njih sam došao do zaključka koje bi varijable bile važne prilikom donošenja odluke o dodjeli kreditne kartice. Varijable ću opisati u tablici koja slijedi u nastavku.

Tablica 1. Opis korištenih varijabli

Varijabla	Opis	Moguće vrijednosti
Osoba	Identifikacija osobe	On (n ={1..500})
Spol	Spol osobe	M – muški Ž - ženski
Godine	Godine života	N = {18..77}
Status	Status osobe na tržištu rada	S – student/ica Z – zaposlen/a N – nezaposlen/a U – umirovljenik/ca
Status zaposlenja	Vremenski period na koji je zasnovan radni odnos	O – određeno N – neodređeno
Staž	Godine radnog staža	N = {0..50}
Obrazovanje	Razina završenog stupnja obrazovanja	NKV - nekvalificirani radnik NSS - niža stručna sprema KV - kvslificirani radnik VKV - visoko kvalificirani radnik SSS - srednja stručna sprema VŠS - viša stručna sprema VSS - visoka stručna sprema DR SC - doktor znanosti MR - magistar
Neto	Prosjek zadnjih 6 mjeseci neto plaća	
Izvor sredstava	Opisuje učestalost prihoda.	R - redovan o - ostalo
Ostali neto	Prosjek ostalih primitaka unazad 6 mjeseci	
Dodjela	Indikator da li je klijent zadovoljio uvjete dodjele kreditne kartice	0 – ne 1 – da

Kada sam odredio varijable, kreirao sam excel dokument te varijable postavio u stupce zatim ručno ispunjavao retke pazeći da ne narušim određeni integritet podataka kako bi podaci bili što realniji, npr. da osoba koja ima 18 godina ne može imati 25 godina radnog staža te biti umirovljenik. Odredio sam da ću ispuniti 500 redaka smatrajući da je to dovoljno za potrebe rada. Nakon ispunjavanja redaka pristupio sam ispunjavanju posljednje varijable za svaki redak. Ta varijabla određuje da li je klijent zadovoljio uvjete dodjele kreditne kartice. U nastavku prikazujem primjer osobe koja zadovoljava i osobe koja ne zadovoljava uvjete koje sam procijenio da bi osoba trebala zadovoljavati.

Osoba	Spol	Godine	Status	Status zaposlenja	Staž	Obrazovanje	Neto	Izvor sredstava	Ostali neto	Dodjela
o1	m	27	z	n	4	VSS	12403	r	0	1
o2	m	27	n		9	SSS		o	2500	0

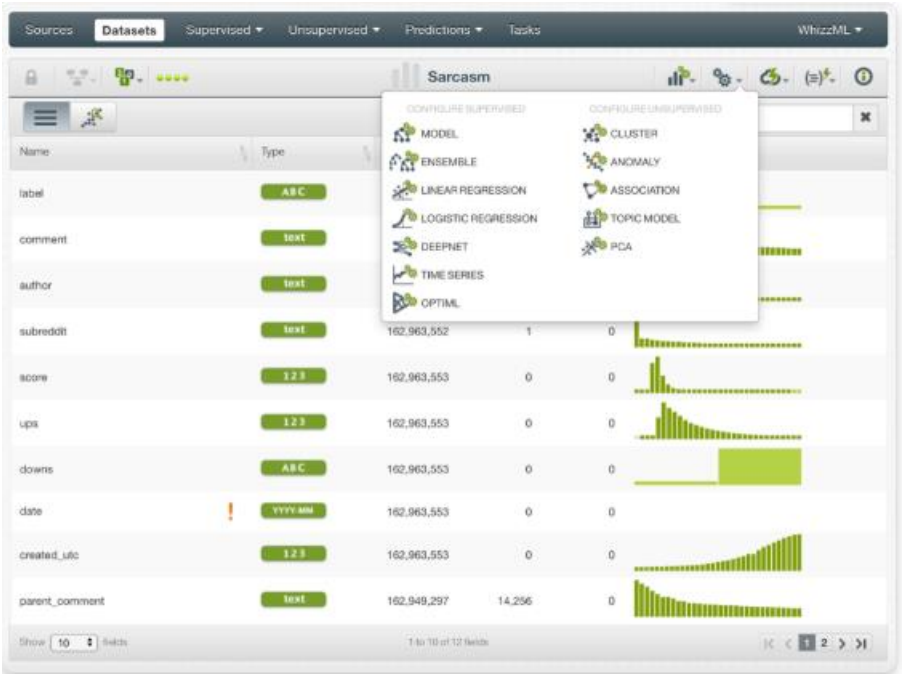
Slika 3. Primjer zapisa iz skupa podataka

Na slici vidimo da o1 zadovoljava uvjete. Osoba je zaposlena na neodređen period sa stažem od 4 godine, visoke stručne spreme te redovnim prosječnim neto primitkom od 12403 kn. Osoba već nekoliko godina radi i sa svojim neto primitkom ima veliku vjerojatnost da može pokrivati troškove nastale na kreditnoj kartici uz ostale životne troškove. Dok o2 istih godina, no većeg radnog staža je trenutno nezaposlena i neto primitak koji ostvaruje nije redovit i teško bi sa njime mogla pokriti troškove kreditne kartice, npr. trošak od petstotinjak kuna bi bio 20% iznosa prosječnog neto primitka koji nije redovan, a i s njime bi osoba vrlo vjerojatno trebala pokriti i još neke prosječne životne troškove nakon čijeg podmirivanja bi se raspoloživi iznos za podmirivanje troškova kreditne kartice znatno smanjio.

Kako se ovdje radi o procjeni da li klijent zadovoljava uvjete za kreditnu karticu na temelju određenih podataka, uvijek postoje i specifične situacije van okvira koje bi se trebale razmotriti izvan ovoga modela te zatim donijeti odluku.

3.3. Softverski alat za primjenu metode stabla odlučivanja

Kao softverski alat za primjenu metode stabla odlučivanja kroz konzultacije sa mentoricom odabrao sam alat dostupan na internetu BigML. Alat je intuitivan i besplatan za korištenje u edukacijske svrhe uz određena ograničenja, no ta ograničenja ne sputavaju izradu modela potrebnog za rad.



Slika 4. Prikaz alata BigML

Kako navode na svojim stranicama, BigML [5], BigMLje platforma za strojno učenje koja se može koristiti za rješavanje zadatka vezanih za klasifikaciju, regresiju, predviđanje vremenskih serija, klaster analize, otkrivanje anomalija, otkrivanje asocijacije i modeliranje tema. Alat pruža selekciju algoritama strojnog učenja koji su dokazani da rješavaju probleme iz stvarnog života te olakšava neograničene aplikacije predviđanja u djelatnostima kao što su zrakoplovstvo, automobilska industrija, energija, zabava, financijske usluge, hrana, zdravstvena zaštita, IoT, farmacija, transport, telekomunikacije i još mnogo toga. Platforma ima mogućnost integracije u ostale sustave.

Algoritam BigML-a, prema BigML [6], je u osnovi nadahnut modelima klasifikacijskog i regresijskog stabla odlučivanja Lea Breiman-a. Koristi mnogo vlastitih komponenti kako bi ga učinili mnogo skalabilnijim i sposobnim da se bavi s više vrsta podataka odjednom. BigML omogućuje stvaranje modela s numeričkim, kategoričkim, datum-vremenskim i tekstualnim poljima.

Osnove:

Za klasifikacijske modele BigML radi podjele na temelju dobivenih informacija. Za regresijske modele podijele smanjuju prosječnu kvadratnu pogrešku (ili smanjenje varijance). Kada generira bodove kandidata za podjelu vezano za numerička polja, BigML ne razmatra svaku moguću podjelu, umjesto toga BigML koristi histograme za odabir kandidata za podjelu (obično 32 kandidata po podijeli).

Strategije za podatke koji nedostaju:

Kada se tijekom treninga nađu podaci koji nedostaju u poljima za unos, postoje dva moguća pristupa: ili ignoriranje nedostajućih slučajeva prilikom generiranja rascjepa kandidata ili eksplicitno uključivanje (npr., godine> 30 ili nedostaju godine).

Kada naiđete na podatke koji nedostaju u vrijeme predviđanja postoje i dvije strategije: s jedne strane BigML generira predviđanje na čvoru stabla čiji se rasplet susreće s nedostajućom vrijednošću (ignorirajući moguću djecu čvora) ili s druge strane BigML ocjenjuje djecu čvora i kombiniraju predviđanja iz oba dijela (slično algoritmu C4.5 i ponekad se naziva "distribucija na temelju imputacije").

Kriteriji zaustavljanja:

Modeli stabla BigML-ovih odluka prestaju rasti kada su iscrpili svoju sposobnost boljeg uklapanja u podatke o treningu, kada su dosegli korisnički definiranu granicu dubine ili kad su dosegli ograničenje čvora definiranu od strane korisnika. Kada koristi ograničenje čvora, BigML bira čvorove koji će se proširiti odabirom onih koji najviše smanjuju grešku u treningu.

Obrezivanje:

BigML nudi statističko obrezivanje kako bi se ublažilo prekomjerno prilagođavanje kad god se trenira model stabala odluka. Orezivanje se temelji na statističkim procjenama pouzdanosti i slično je orezivanju u algoritmu C4.5.

Vrednovanje:

Modeli stabla BigML-ovih odluka također prihvaćaju ponderirane podatke o treningu. Za regresijske modele svaki se primjerak može mjeriti pojedinačno, no podaci se mogu mjeriti po razredima kod rada s klasifikacijskim modelima.

Rad sa ovim alatom će biti detaljnije opisan u poglavlju gdje se razrađuje tema.

4. Analiza prethodnih istraživanja

U ovom poglavlju detaljnije ću opisati svoja zapažanja prilikom analize stručnih članaka, pokušati prenijeti neke ideje autora te se koncentrirati na dijelove koji su bliži temi moga rada.

4.1. Primjena stabla odlučivanja u kreditnom skoringu

Autorica doc.dr.sc. Ljiljanka Kvesić sa Fakulteta prirodoslovno-matematičkih i odgojnih znanosti sveučilišta u Mostaru, BiH.

U uvodu članka detaljno je opisan pojam kreditnog skoringa koji je, kako navodi autorica, sustav dodjeljivanja bodova zajmotražitelju pri čemu se primjenjuje statistička analiza ključnih karakteristika zajmotražitelja s ciljem numeričke kvantifikacije rizika koja pokazuje vjerojatnost da će klijent doći u status neispunjavanja ugovorene obveze. Takav sustav dodjeljuje jednu kvantitativnu mjeru, koja se naziva skor, potencijalnom klijentu predstavljajući njegovo buduće ponašanje u otplati dodijeljenog kredita. Takav skor se uspoređuje sa najniže prihvatljivim brojem bodova, definirano kreditnom politikom banke, na temelju čega se kredit odobrava ili ne odobrava. Ovakav sustav nema direktne poveznice sa mojom temom, ali daje nekakvu ideju da također na temelju nekih karakteristika tražitelja kreditne kartice treba biti donesen indikator na temelju kojeg će se znati da li je tražitelj kartice zadovoljio uvjete da mu se odobri izdavanje i korištenje kreditne kartice.

Dalje u uvodu, autorica opisuje skoring modele pa navodi da postoje aplikativni i bihevioralni skoring modeli. Aplikativni se koriste u slučaju novih klijenata, a bihevioralni u slučaju postojećih klijenata. Razlika je, uz još neke o kojima neću ovom prilikom, u tome da bihevioralni skoring modeli upotrebljavaju informacije o ponašanju klijenta pri otplati kredita u prošlom periodu, periodu promatranja, a ujedno se mogu koristiti i podaci prikupljeni kod podnošenja zahtjeva za kredit. Dok se kod aplikativnih skoring modela koriste samo podaci prikupljeni prilikom podnošenja zahtjeva. Kako se u mojoj temi radi o izdavanju kreditne kartice novim korisnicima, u ovom članku sam dobio smjernicu da se moram usredotočiti na promatranje aplikativnih skoring modela.

Autorica [1] je kao cilj njezinog rada označila razvoj kreditnog skoring modela, utemeljenog na stablu odlučivanja, koji bi bio primjenjiv u poslovanju hrvatskih financijskih institucija te na taj način popunio prazninu u teoriji i praksi. Suprotno od mog rada, ovdje je koncentracija na bihevioralnom skoring modelu i istraživanju kojim bi se otkrilo koje su karakteristike ključne za predikciju statusa ne ispunjavanja obaveza. Kako sam već naveo da

bihevioralni skoring model koristi i podatke koji su prikupljeni prilikom ispunjenja zahtjeva uočio sam da bio ovaj dio rada autorice mogao biti od pomoći prilikom izrade moga rada.

Autorica [1] u članku navodi i prethodna istraživanja gdje se predstavljaju mogućnosti korištenja stabla odlučivanja pri procjeni kreditne sposobnosti zajmotražitelja pa ću ih ja uz malo detalja i navesti:

1) Joos et al. su koristeći opsežnu bazu podataka jedne od najvećih belgijskih banaka usporedili rezultate logit modela i stabla odlučivanja te se prema autorima globalna točnost logit modela pokazala veća u slučaju cjelovite sheme skupa podataka dok su za kvalitativne podatke i kraće sheme, važnija bila stabla odlučivanja.

2) Zurada i Zurada su istraživanjem podataka, financijskih obilježja svakog klijenta, o kreditima koje je odobrila jedna financijska institucija primijenili tehnike stabla odlučivanja, neuronske mreže i logit regresiju te model koji kombinira sve tri metode. Kako bi usporedili učinkovitost svakog korištenog modela razvili su dva scenarija te zaključuju da su najbolje rezultate klasifikacije dobili primjenom neuronskih mreža i modela koji je kombinirao tri analizirane tehnike.

3) Xiao, Zhao i Fei ispitali su performanse različitih modela kreditnog skoringa i pripadajuće troškove kreditnog rizika. Na temelju podataka njemačkih i australijskih zajmotražitelja definirali su 24 varijable. Za klasifikacijske algoritme koristili su linearnu diskriminacijsku analizu, logističku regresiju, neuronske mreže, k-najbližih susjeda, metodu potpornih vektora, klasifikacijska i regresijska stabla te metodu MARS(Multivariate Adaptive Regression Splines). Na temelju dobivenih rezultata zaključili su da metoda potpornih vektora, MARS, logistička regresija i neuronske mreže imaju vrlo dobre performanse, najviše prilikom klasifikacije dobih klijenata, dok su se linearna diskriminacija, klasifikacijska i regresijska stabla pokazala značajno preciznijim prilikom identifikacije loših zajmotražitelja.

4) Satchidananda i Sima su koristili podatke dvaju indijskih banaka kako bi usporedili učinkovitost stabala odlučivanja i logističke regresije prilikom predikcije dobrih i loših zajmotražitelja. Fokusirali su se na kredite vezane uz poljoprivrednu proizvodnju. Zaključak je da stabla odlučivanja bolje klasificiraju dobre i loše klijente nego model logističke regresije.

Autorica navodi još 4 istraživanja kod kojih su se metode stabla odlučivanja pokazale najučinkovitijim prilikom klasifikacije dobrih i loših zajmotražitelja.

U radu, autorica se odlučuje za korištenje algoritma klasifikacijskog stabla kod kojeg se skup odgovora iz zahtjeva za kredit dijeli na dva podskupa gdje je razlika u prosječnom riziku neispunjenja obveza između dva podskupa što je moguće veća. Kako je pretpostavka primjene stabla odlučivanja posjedovanje baze podataka kandidata, koji su opisani s n atributa, autorica navodi da je uzorak dobiven istraživanjem iznosio 200 klijenata jedne poslovne banke od kojih je bilo 100 dobrih(ne rizični) i 100 loših(rizični).

Kao cilj modela označen je pronalazak atributa koji najbolje razdvaja uzorak dobrih klijenata od uzorka loših klijenata. Algoritam počinje čvorom koji sadrži uzorke dobrih i loših klijenata, nakon čega se pronalaze svi ishodi s ciljem koji je već naveden, a to je pronalazak najkorisnijeg atributa i odgovarajuće granične vrijednosti koja najbolje vrši razdvajanje uzoraka klijenata.

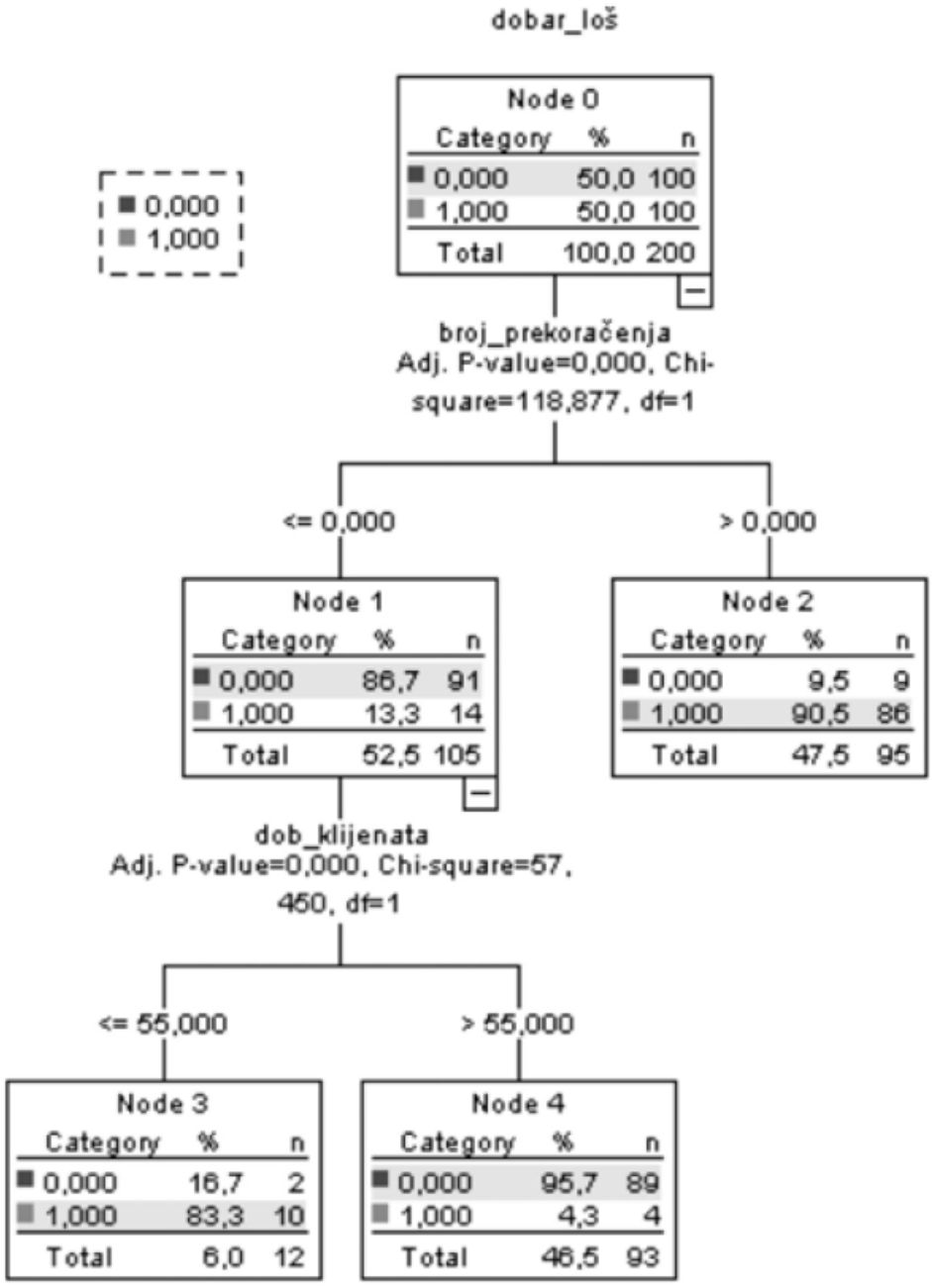
Konstruiranje stabla odlučivanja opisuje u tri koraka:

1) izgradnja logičkog modela
2) računanje očekivanih vrijednosti odluka postupkom računanja unatrag
3) pronalaženje optimalnog puta postupkom računanja prema naprijed

Za testiranje svakog razdvajanja i mjerenje homogenosti podataka koristila je Kolmogorov-Smirnovljevu statistiku, indeks diverzifikacije, Gini-jev koeficijent i indeks entropije.

Navedene su i prednosti stabla odlučivanja: mogućnost generiranja razumljivih modela, sa čime se u potpunosti slažem, relativno mali zahtjevi za računalnim izvorima, također se slažem pošto algoritam stabla odlučivanja nije odveć zahtjevan za razliku od nekih drugih algoritama, jasna važnost pojedinih atributa za konkretni problem, isto se slažem jer dobije se jasna slika o važnosti i povezanosti atributa sa problemom i na kraju kao prednost je navedena i široka dostupnost softverskih alata i rješenja u što sam se uvjerio prilikom traženja alata koji bi koristio u svrhu izrade svog rada gdje sam naišao na pregršt besplatnih i programskih rješenja koja se plaćaju. Uz prednosti naveden je i jedan nedostatak kao nestabilnost stabala odlučivanja jer kod malih fluktuacija u uzorku podataka mogu se dogoditi velike varijacije u dodijeljenim klasifikacijama.

Za konstrukciju stabla odlučivanja korišten je hi-kvadrat automatizirana interakcijska detekcija (CHAID algoritam). Kod navedenog algoritma u svakom se koraku izabire nezavisna varijabla koja ima najjaču interakciju sa zavisnom varijablom. Ako se s obzirom na zavisnu varijablu značajno ne razlikuju, modaliteti svakog prediktora se spajaju. Iscrpni CHAID algoritam predstavlja modifikaciju izvorne metode u kojoj se ispituju sva moguća razdvajanja svakog prediktora.



Slika 5. Stablo odlučivanja dobiveno CHAID algoritmom pomoću SPSS statističkog paketa [1]

Kao zavisna varijabla odabrana je varijabla definirana kao dobar/loš klijent, a kao nezavisne (prediktorske) definirane su spol, dob, status klijenta, minimalno dozvoljeno prekoračenje tijekom 6 mjeseci, prosječan iznos svih plaćanja čekovima, prosječan iznos plaćanja karticom, prosječno stanje na računu za 6 mjeseci, prosječan iznos sredstava, koje klijent još smije potrošiti, broj nedozvoljenih prekoračenja.

Zaključak na temelju dobivenog stabla je da, među klijentima koji su se najmanje jedanput nalazili u nedozvoljenom prekoračenju bilo 90,5% loših klijenata. Među klijentima koji nisu bili niti jednom u nedozvoljenom prekoračenju, a imaju 55 ili manje godina bilo je 83,3% loših klijenata, dok je među klijentima koji nisu bili niti jednomu nedozvoljenom prekoračenju, a imaju više od 55 godina, bilo 95,7% dobrih klijenata. Ispravno je klasificirano 89 dobrih i 96 loših klijenata. Dakle, 11 dobrih i 4 loša klijenta nisu bila ispravno klasificirana. Ukupno je 92,5% klijenata korektno klasificirano pa autorica model smatra zadovoljavajućim.

Ovaj rad autorice me najviše usmjerio i imao najveći utjecaj prilikom izrade moga rada. Uz neke razlike bila je prisutna doza sličnosti koju sam iskoristio kako bi se kroz ogromnu teoriju usmjerio u pravome smjeru.

4.2. Procjena kreditnog rizika pomoću statističkog i strojnog učenja: Osnovne metodologije i aplikacije za modeliranje rizika

Originalni naslov članka na engleskom jeziku je „Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications“, autori su J. Galindo (sa odjela za ekonomiju sveučilišta Harvard iz grada Cambridge koji se nalazi u saveznoj državi Massachusetts, SAD) i P. Tamayo (iz tvrtke Thinking Machines Corp., iz grada Burlington, također savezna država Massachusetts, SAD).

Prema autorima [7], procjena rizika financijskih posrednika područje je novih interesa u 2000. godini, zbog financijske krize 1980-ih i 90-ih koja najviše pogađa SAD, Meksiko, nordijske zemlje i Japan. Točna procjena rizika i njegova upotreba u korporacijama ili globalni modeli financijskog rizika mogli bi se prenijeti u učinkovitije korištenje resursa. Kako bi se postigao taj cilj potrebno je pronaći precizne prediktore pojedinačnog rizika portfelja institucija. U tom kontekstu radi se usporedna analiza različitih statističkih modela kao što su statistička regresija, stabla odlučivanja (CART), neuronske mreže i *k*-najbližih-susjeda koristeći isti skup podataka i metode modeliranja strojnog učenja klasifikacije na skupu podataka o hipotekarnom zajmovima, dobivenim od jedne velike komercijalne banke, s motivacijom razumijevanja njihovih ograničenja i potencijala. Autori uvode specifičnu metodologiju modeliranja baziranu na proučavanju krivulja pogreške. Korištenjem vrhunskih tehnika modeliranja izgradili su više od 9.000 modela kao dio studije. Rezultati pokazuju da modeli s CART stablima odlučivanja pružaju najbolju procjenu za zadane vrijednosti s prosječnom stopom pogreške 8,31% za trenažni uzorak od 2000 zapisa. Kao rezultat analize krivulje pogreške za ovaj model zaključuju da, ukoliko je na raspolaganju više podataka, otprilike 22.000 zapisa, može se postići potencijalna stopa pogreške 7,32%. Pošto se ova metoda pokazala kao najbolja, a ima poveznice sa temom moga rada, kroz ovaj osvrt predstavit ću je malo detaljnije.

Za CART (eng. classification and regression tree, hrv. klasifikacijsko i regresijsko stablo) autori navode da se radi o moćnom ne parametarskom modelu koji daje precizne procjene i lako razumljiva pravila kojima je model okarakteriziran. Jedan je od dobrih predstavnika stabala odluke uz C5.0, CHAID koji je korišten u prethodnom članku, NewID, Ca15 i ostale modele stabla odluke. Ističu da je dobra karakteristika ovoga modela njegova transparentnost jer može biti prikazan kao niz pravila napisanih riječima što ga čini idealnim za ekonomske i financijske aplikacije. U tablici su predstavljeni rezultati početnih istraživanja

vezanih za posljedice promjene parametara. Veličinu klasifikacijskog i regresijskog stabla su kontrolirali mijenjajući parametar gustoće koji predstavlja minimalni postotak zapisa u bilo kojoj klasi koja zahtjeva razdvajanje u bilo kojem čvoru stabla.

Tablica 2. Usporedba preciznosti i vremenske razlike u klasifikacijsko regresijskog modela stabla[7]

Gustoća	Veličina stabla(najbolje stablo)	Veličina stabla (najveće stablo)	Greška testa(%)	Vrijeme(sekunda)
0.2	25	25	10.5	13
0.15	25	25	10.5	13
0.1	35	41	9.8	13
0.05	39	45	7.5	14
0.0025	81	89	7	17
0.01	77	121	6.9	20
0.005	109	189	6.5	24
0	161	299	6.7	27

Promjenom parametra gustoće autori su mogli promatrati odnos između parametara preciznosti, veličine modela i vremena potrebnog za izradu modela. U tablici pod rednim brojem 1 vidi se da smanjenjem gustoće povećavaju se vrijeme potrebno za izradu modela i preciznost samog modela. Na početku je gustoća velika kako bi se dobio početni model koji se kasnijim ugađanjima precizira. Početni model napravljen je sa gustoćom 0.05 kako bi se saznala veličina stabla i vrijeme potrebno za izradu. Za funkciju nečistoće korišten je Gini indeks. Varijabla veličina stabla (najbolje stablo) označava veličinu podstabla cjelokupnog klasifikacijskog i regresijskog stabla koji ima najmanji postotak greške nad testnim skupom podataka. Takvo drvo nastaje procesom „obrezivanja“ kojim se kreira skup podstabala koji nastaju eliminiranjem grupa grana. Eliminiranje grupa grana se izvršava uzimajući u obzir zamršenost i razliku u postotcima pogreške originalnog klasifikacijskog i regresijskog algoritma. Autori navode da taj proces za stablo odlučivanja omogućuje praktične mjere za veličinu modela i kontrolu njegova kapaciteta.

Analizi krivulje učenja pristupaju sa modelima koji imaju 20, 40, 80, 100, 120, 200, 300 i 400 čvorova te skupovima podataka sa 64, 128, 256, 500, 750, 1000, 1250, 1500, 1750 i 2000 zapisa. Za analizu je korišteno 2400 modela. Kako raste broj zapisa tako se smanjuje greška generalizacije. Kao najbolji model pokazao se onaj sa 120 čvorova koji ima prosječni postotak pogreške 8,31% na skupu podataka od 2000 zapisa.

Neuronske mreže osigurale su drugi najbolji rezultati s prosječnom pogreškom od 11,00%. K-Najbliži susjed algoritam imao je prosječnu stopu pogreške od 14,95%. Ti su rezultati nadmašili standardni algoritam Probit koji je dostigao prosječnu stopu pogreške 15,13%. Počinje svijest u svijetu da je potrebno koristiti ove vrste preciznih prediktivnih modela u institucijskim i globalnim modelima rizika.

Članak je iscrpno potkrepljen velikom količinom radova drugih autora koji pokrivaju određene teme koje se spominju kroz rad. Također, podaci potrebni za procjene modela su odlično vizualno prikazani i analizirani što olakšava njihovo razumijevanje i usporedbu. Vidimo da se 2000. godine budi svijest o korištenju prediktivnih modela i njihovoj važnosti u prevenciji rizika, no slična situacija se ponovila sredinom i krajem početka drugog milenija pa se postavlja pitanje da li se modeli nisu koristili, nisu bili dobro/dovoljno postavljeni/razvijeni ili su jednostavno bili zanemareni iz nekih sebičnih razloga.

4.3. Potrošački modeli kreditnog rizika putem algoritama strojnog učenja

Originalni naslov članka na engleskom jeziku je „Consumer credit-risk models via machine-learning algorithms“ grupe autora Amir E. Khandani, Adlar J. Kim i Andrew W. Lo sa MIT škole menadžmenta i laboratorija za financijski inžinjering.

Autori [9] su primijenili tehnike strojnog učenja kako bi izgradili nelinearne bez parametarske modele predviđanja kreditnog rizika potrošača. Kombinacijom podataka o transakcijama klijenata i podatke iz kreditnih institucija od siječnja 2005. do travnja 2009. za uzorak klijenata većih komercijalnih banaka, bili su u mogućnosti konstruirati prognoze izvan uzorka koje značajno poboljšavaju stope klasifikacije prestupnika i nepodmirenih obaveza vlasnika kreditnih kartica, sa R2 linearne regresije predviđene/realizirane delinkvencije od 85%. Korištenje konzervativnih pretpostavki za troškove i koristi rezanja kreditnih mogućnosti na temelju predviđanja strojnog učenja procijenili su da se trošak uštede kreće od 6% do 25% ukupnih gubitaka. Navode da uzorci vremenskih nizova procijenjene stope delinkvencije njihovog modela tijekom nedavne financijske krize ukazuju na to da agregirane analitike kreditnog rizika korisnika, mogu imati važnu primjenu u prognoziranju sustavnog rizika. Autori ističu da su prepoznali uzorak ponašanja klijenta koji prethodi delikventnom ponašanju.

Odlučili su se ovaj izazov formulirati kao problem koji će se riješiti korištenjem nadziranog učenja. Kod nadziranog učenja predstavljaju se parovi ulaza i izlaza u obliku (ulaz,izlaz) = (x,y). Ako je y diskretna/nebrojčana vrijednost radi se o klasifikaciji, a ako je y kontinuirana/brojčana vrijednost radi se o regresiji. Izlaz je u modelu autora kontinuirana vrijednost koja se kreće između 0 i 1 i predstavlja vjerojatnost da će se pojaviti delikventno ponašanje na pojedinom računu kreditne kartice. Model se temelji na algoritmu klasifikacijskih i regresijskih stabala odlučivanja.

Autori [9] su mišljenja da popularnost modela klasifikacijskih i regresijskih stabala odlučivanja proizlazi iz činjenice da prelazi ograničenja standardnih model kao što su logit i probit u koje je ovisan varijabla prisiljena uklopiti se u jedan linearni model za cijeli ulazni skup. Navode njegovu mogućnost da otkrije nelinearne interakcije između ulaznih varijabli što značajno povećava tipove relacija koji mogu biti obuhvaćeni kao i broj zavisnih varijabli koje se mogu koristiti. Također, model predstavlja lako razumljiva pravila odluke čija je logika jasno prikazana u stablu. Kao što je navedeno i u prethodnom članku, takav model je pogodan za korištenje u bankarskom sektoru jer nije model „crne kutije“ na koji se gleda sa dozom sumnje i skepticizma. Autori također spominju proces obrezivanja koji se temelji na Gini indeksu i koji

se koristi za „ograničavanje“ rasta stabla. Navode i formulu Gini indeksa gdje je τ krajnji čvor, a $P_\tau(k)$ udio podataka za trening dodjeljenih klasi k na navedenom krajnjem čvoru:

$$G(\tau) \equiv \sum_{k=1}^{K} P_\tau(k)(1 - P_\tau(k))$$

A kriterij „obrezivanja“ je predočen formulom, gdje |T| predstavlja broj krajnjih čvorova u modelu T, a λ predstavlja regulatorski parametar odabran unakrsnom validacijom:

$$C(T) \equiv \sum_{\tau=1}^{|T|} G(\tau) + \lambda|T|$$

Kada kriterij „obrezivanja“ dođe do minimuma, algoritam prestaje sa širenjem stabla.

Izlaz modela je kontinuirana varijabla, koja se može predstaviti kao procjena vjerojatnosti da će se na računu pojaviti delikventno ponašanje koje će trajati 90 ili više dana. Razlika između procjene računa koji su postali delikventni i onih koji nisu je jedna mjera uspješnosti modela, u slučaju da su računi koji su postali delikventni imali istu procjenu kao oni koji nisu, prognoze nemaju vrijednost.

Autori uspoređuju prosječnu procjenu modela postojećih klijenata čiji je račun postao delikventan sa prosječnom procjenom postojećih klijenata kojima račun nije postao delikventan. Njihova očekivanja su da rezultat modela neće biti impresivan, no vrijednosti su pokazale da je model nevjerojatno moćan i može jasno razdvojiti dvije populacije. Kao primjer navode razdoblje iz travnja 2008. godine gdje je prosječna prognoza modela ,postojećih klijenata koji nisu postali delikventni između svibnja i srpnja 2008. godine , 0.7 dok je prosječna procjena potencijalnih delikvenata bila 10.3, a stupanj odvojenosti je dosljedan tijekom 10 razdoblja ocijenjivanja. To potvrđuje da model ima zadovoljavajuću preciznost usprkos tome što je rađen na malom skupu podataka, a autori se nadaju da će njihov rezultat potaknuti izradu moćnih modela ponašanja korisnika uz pomoć strojnog učenja i određena ugađanja i veći skup podataka.

5. Metoda stabla odlučivanja za određivanje rizika davanja kreditne kartice

5.1. Opis problema

Izdavanje kreditnih kartica čini dobar dio poslovanja banaka u Republici Hrvatskoj. To je popularan i učinkovit način kratkoročnog kreditiranja sa ne pretjerano velikim pojedinim iznosima. Izdavanjem takvih kartica banka se izlaže potencijalnom riziku i da bi se zaštitila od gubitaka mora biti u mogućnosti dobro procijeniti rizik te donijeti odluku o kreditiranju sukladno tome.

Popularizacijom metoda strojnog učenja i statističkih modela te razvojem moći računala uz prihvatljivu cijenu sve više se javlja njihovo korištenje u domeni banaka, financijskog i ekonomskog poslovanja. Dosadašnje odluke o kreditiranju bile su bazirane na subjektivnoj ocjeni analitičara koji je morao imati određeno znanje i iskustvo te slijediti niz propisanih pravila, a kako je danas znanje skupo i za sticanje iskustva potrebno je dosta vremena, teži se izradi skoring modela koji su bazirani na statističkim metodama i modelima strojnog učenja.

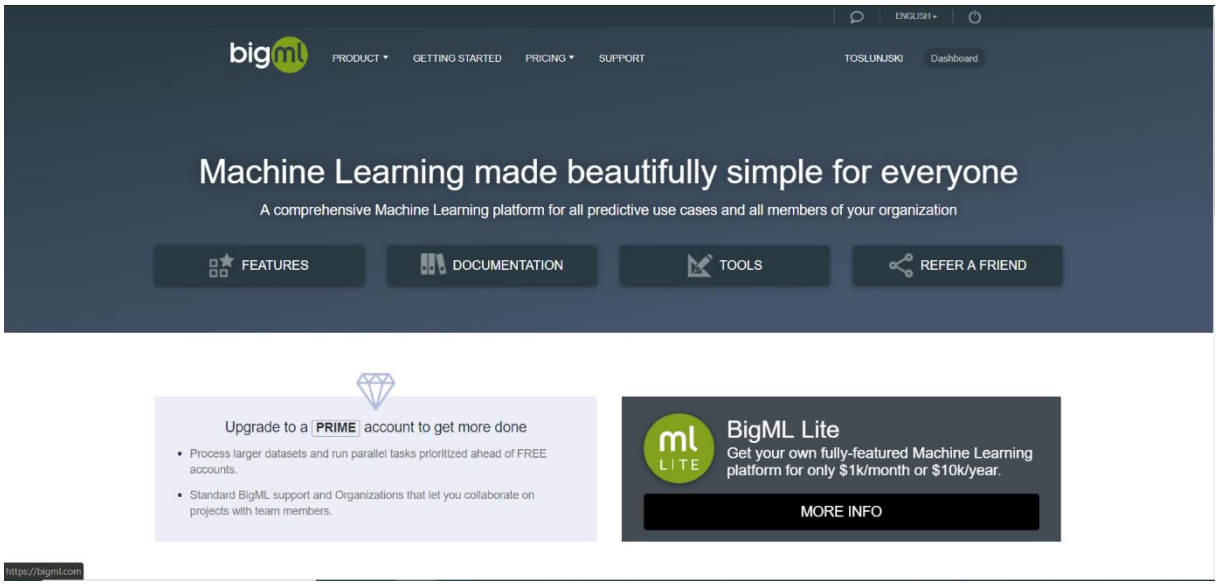
Takav model olakšao bi donošenje odluke analitičaru koji bi samo trebao unijeti određene podatke te bi na temelju njih dobio rezultat uz pomoć kojeg bi lakše, jednostavnije i brže donio odluku. Uz to model bi trebao biti jasan, razumljiv i lako čitljiv kako bi se u slučaju potrebe mogla opravdati donesena odluka.

Kroz analizu stručne literature i konzultacije sa mentoricom donio sam odluku o izradi modela primjenom metode stabla odlučivanja koji bi procjenjivao da li je tražitelj kreditne kartice dobar kandidat, tj. da li mu se zahtjev treba odobriti ili ne.

5.2. Rješavanje problema primjenom metode stabla odlučivanja

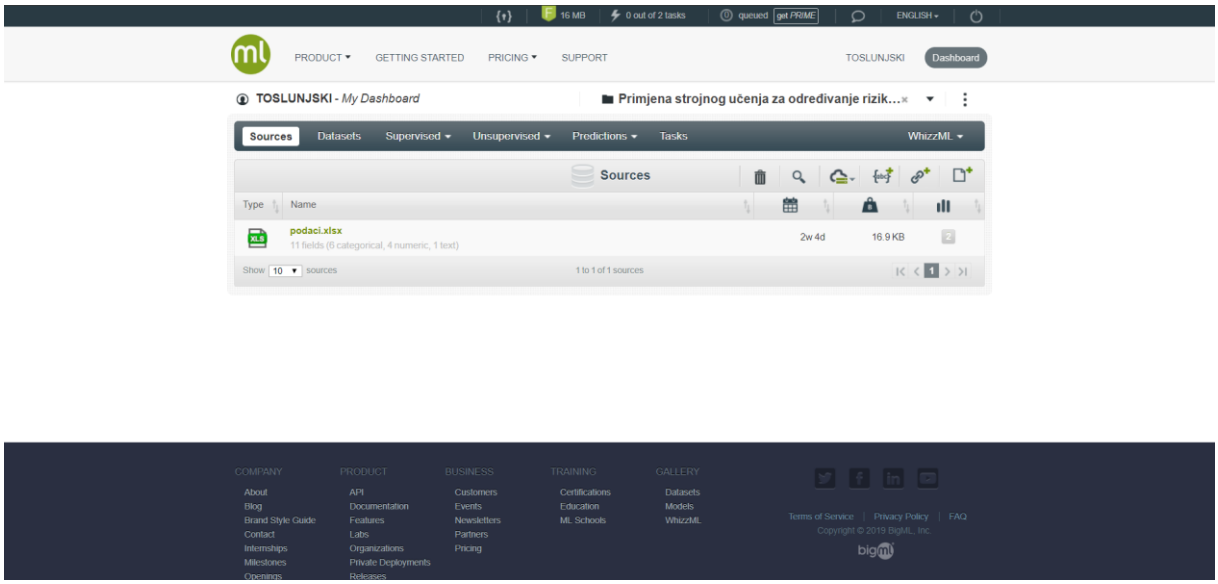
5.2.1. Skup podataka

U alatu BigML potrebno je kreirati izvor podataka (data source), radi se o „sirovim“ podacima na temelju kojih se napravi skup podataka (data set). To se radi tako da nakon prijave u alat odaberemo kontrolnu ploču (dashboard).



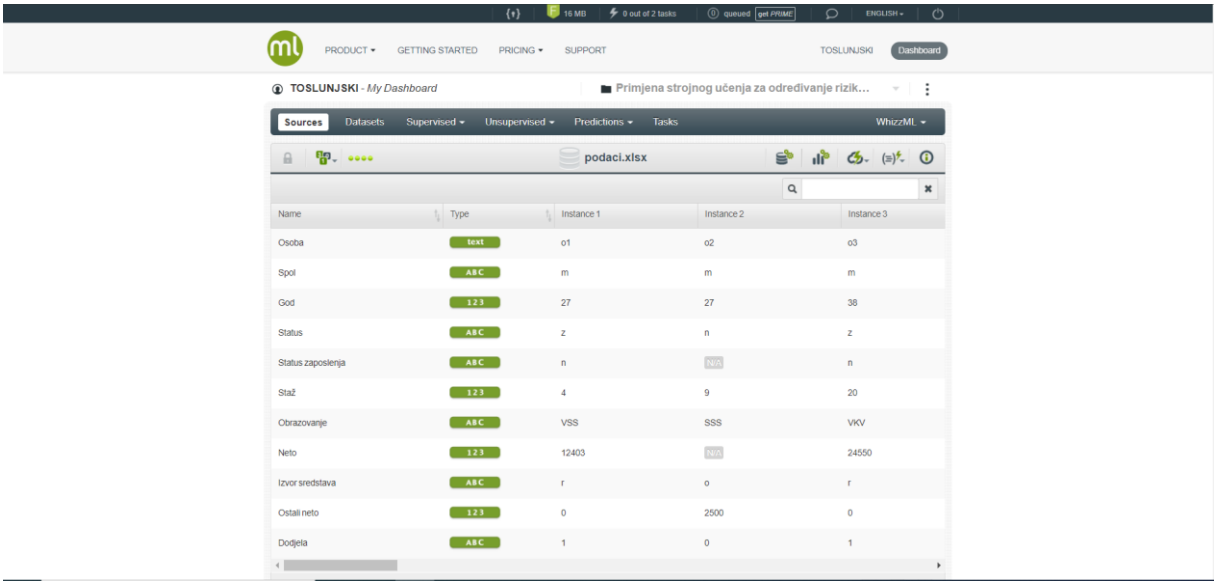
Slika 6. Glavni ekran aplikacije BigML nakon prijave u aplikaciju

Zatim odaberemo karticu „Izvori“ (Sources). Pritiskom na „File“ gumb otvara se opcija uvoza podataka sa lokalnog računala. Odabirem xlsx dokument sa svojim simuliranim podacima te ga uvozim u alat. Alat nudi i dodatne opcije uvoza podataka.



Slika 7. Prikaz kontrolne ploče u aplikaciji BigML sa karticama

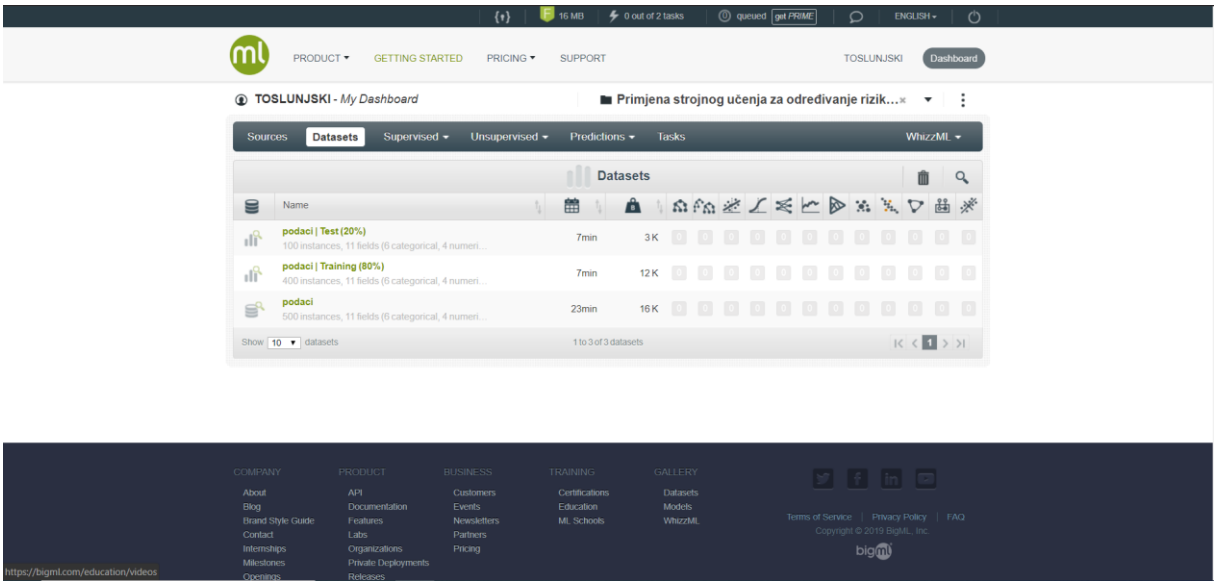
Za pregled izvora podataka potrebno je pozicionirati se na naziv izvora i pritisnuti tipku miša. Na pregledu izvora podataka vidimo popis varijabli, oznake njihovih tipova te same vrijednosti varijabli.



Slika 8. Pregled izvora podataka u aplikaciji BigML

Kada je kreiran izvor podataka sa izvornim podacima, kreirat ću skup podataka. To se radi tako da se pozicionira na ikonu oblaka u kojem se nudi opcija „jednim pritiskom miša do skupa podataka“(eng. 1-click dataset). Na taj način dobio sam skup podataka. Taj skup podataka podijelio sam na dva dijela. Podjela je napravljena metodom nasumičnog razdvajanja. Tako je dobiven skup naziva trening u kojem se nalazi 80% podataka skupa podataka kreiranog na početku. Njega ću koristiti za treniranje prediktivnog modela. Drugi

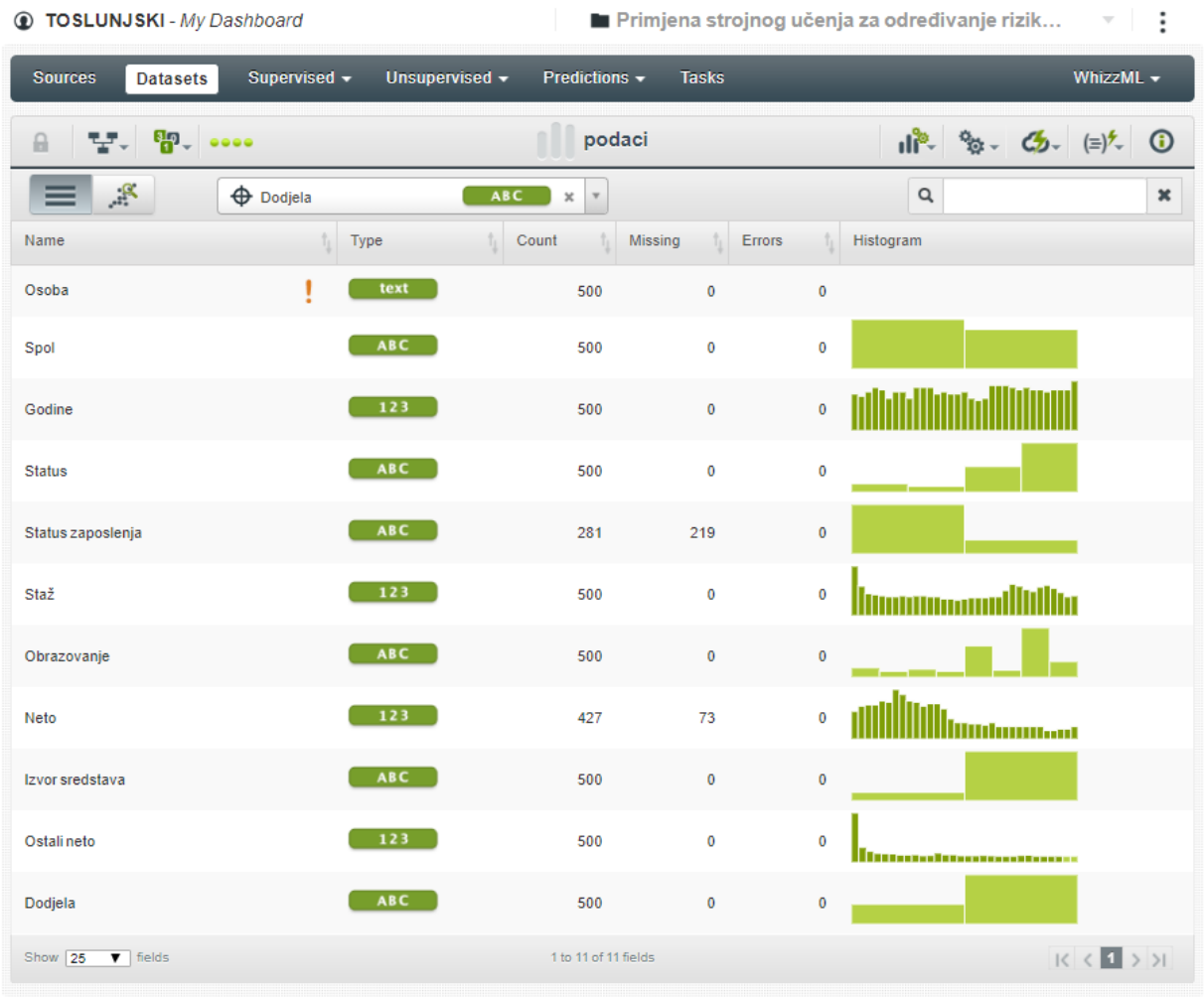
skup, naziva test, sadrži 20% podataka od početno kreiranog skupa koje model za vrijeme treniranja neće vidjeti i njega ću koristiti za evaluaciju kreiranog prediktivnog modela. Sada kada imam potrebne skupove podataka mogu pristupiti izradi modela.



Slika 9. Popis skupova podataka u aplikaciji BigML

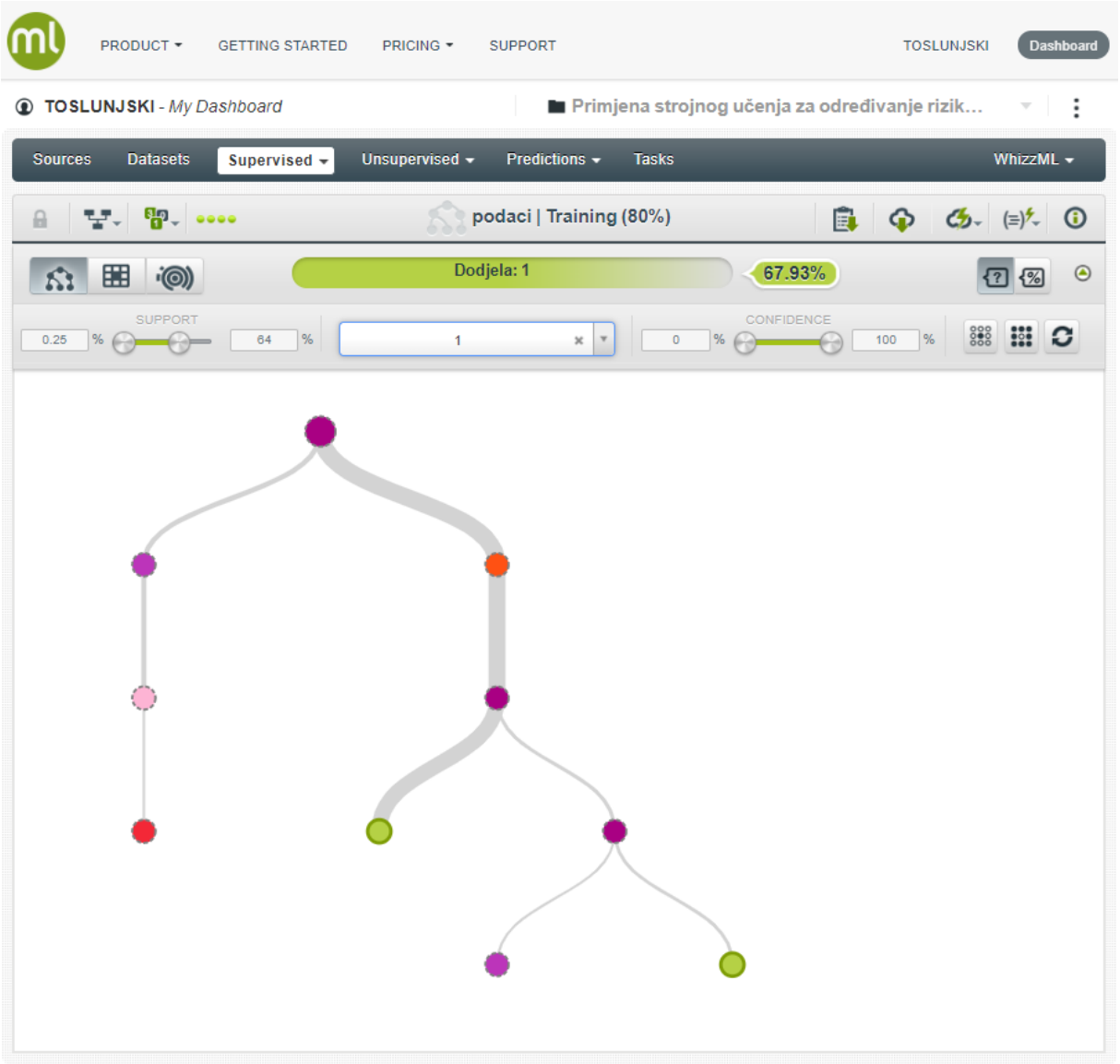
5.2.2. Izrada modela

Model je u aplikaciji BigML vrlo jednostavno izraditi. Preduvjet je kreirani skup podataka. Na kartici „skupovi podataka“ (eng. datasets) potrebno je pozicionirati se na naziv skupa podataka te pritisak miša kako bi se pristupilo detaljima samog skupa. Prvo kreiram prediktivni model na temelju trening podataka pa sukladno tome otvaram trening skup podataka. Pozicioniranjem i pritiskom na ikonu oblaka otvara se izbornik u kojem odabiremo, među ostalim opcijama, opciju „jednim pritiskom miša do nadziranog modela“ (eng. 1 click supervised model).



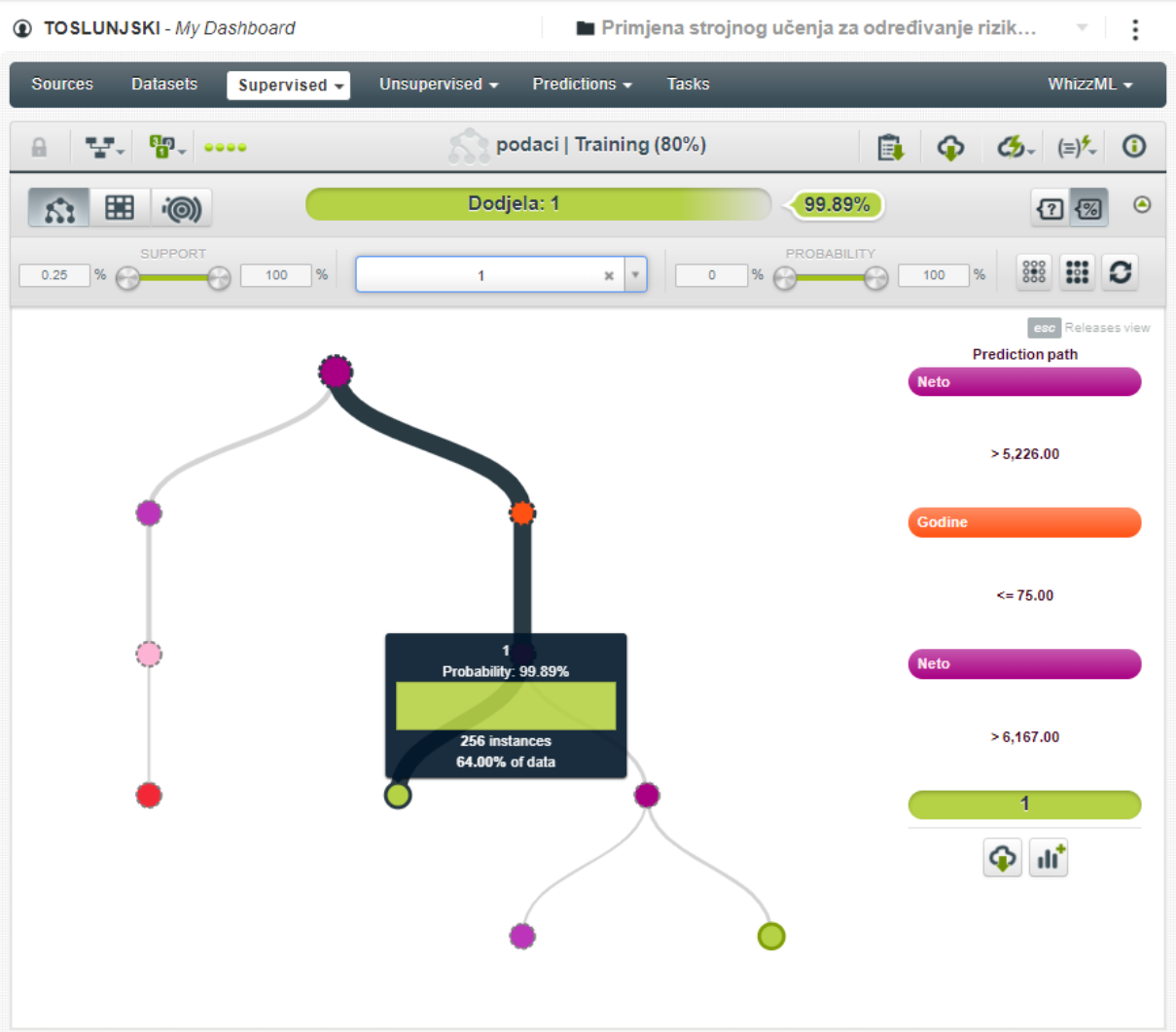
Slika 10. Detalji skupa trening podataka u aplikaciji BigML

Alat automatikom izrađuje model te ga prikazuje. Da li je model klasifikacijski ili regresijski ovisi o tipu podatka ciljane varijable. U mom slučaju ciljana varijabla je „Dodjela“ koja označava da li se dodjeljuje kreditna kartica ili ne. Njen tip podatka je kategorijski stoga će model biti klasifikacijski.



Slika 11. Prikaz prediktivnog modela stabla odlučivanja na temelju trening podataka u aplikaciji BigML

Na slici 11 vidimo dobiven grafički prikaz modela stabla odlučivanja. Stablo ima dvije glavne grane, lijevu i desnu. Što je grana deblje prikazana to je veći broj instanci koje se nalaze u toj grani. Glavne grane izlaze iz korijenskog čvora koji najbolje razdvaja podatke na dva segmenta vezano za ciljanu varijablu. U korijenskom čvoru nalazi se varijabla „Neto“ koja predstavlja prosjek neto plaće unazad 6 mjeseci. Desna grana je deblje prikazana što znači da ona u sebi sadrži najveći broj instanci. Na toj grani nalaze se još čvorovi „Godine“ pa opet „Neto“ kod kojeg imamo još jedno grananje i na kraju ciljana varijabla sa klasom „1“ koja označava dodjelu kreditne kartice.



Slika 12. Detaljniji prikaz čvorova grane stabla odlučivanja u aplikaciji BigML

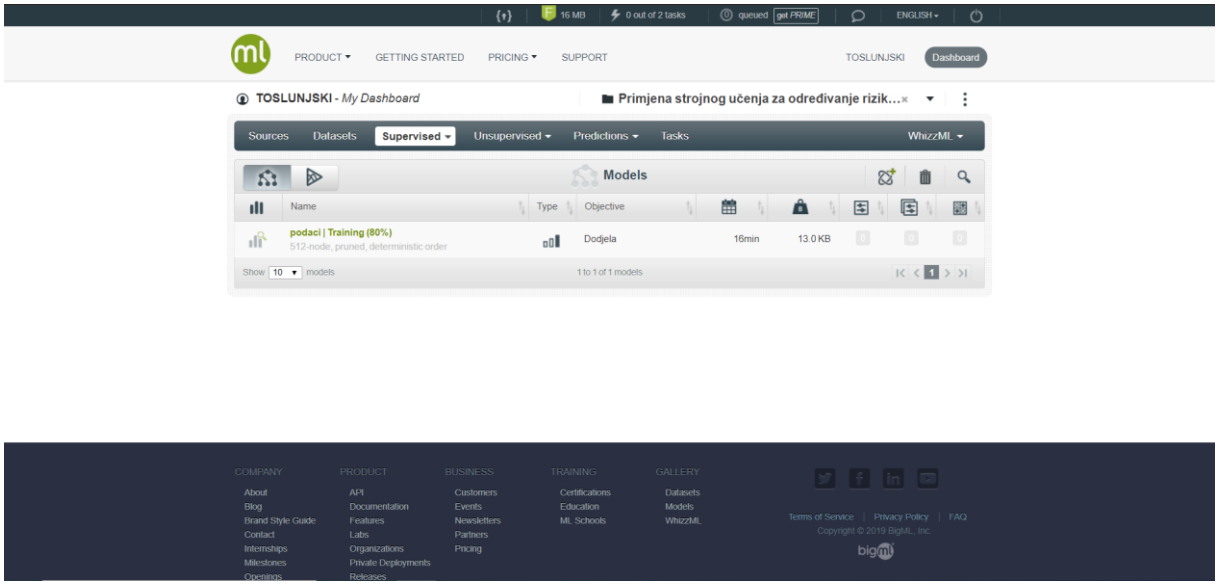
Dakle na detaljnom prikazu sa slike 12 vidimo da je u korijenskom čvoru varijabla „Neto“ te da će sve instance sa vrijednosti varijable većom od 5,226.00 ići granom koja vodi prema mogućim odobrenjima zahtjeva za izdavanje kreditne kartice. Sljedeći čvor na desnoj grani je čvor u kojem se nalazi varijabla „Godine“ iz kojeg će dalje nastaviti instance koje imaju vrijednost varijable manje ili jednako 75.00. Samo napomena da BigML radi jednostavnijeg prikaza ne prikazuje sva grananja, nego samo ona relevantnija. Nakon tog čvora slijedi ponovno čvor sa varijablom „Neto“ gdje se promijenila vrijednost pa će sada direktno do čvora lista sa klasom 1 ići one instance koje imaju neto veći od 6,167.00, dok će ostale ići na daljnje grananje. Po opisanoj grani do čvora lista došlo je 256 instanci, sa vjerojatnošću od 99.89%, što čini 64% podataka.

Prikazujem i objašnjavam samo jednu granu kako bi se shvatila poanta grafičkog prikaza stabla. Kao što sam naveo u grafičkom prikazu nisu prikazane sve grane radi jednostavnosti prikaza, no sa kretanjem po čvorovima stabla može se vidjeti detaljniji prikaz. Ovaj grafički prikaz ujedno prikazuje i pravila koja se kriju iza odluka i tu do izražaja dolazi spominjana transparentnost i lako razumijevanja odluka stabla odlučivanja. Uz grafički prikaz stabla odlučivanja postoje i ostali prikazi koji na druge načine, također dosta transparentno, prikazuju odnos pravila i podataka instanci.

5.2.3. Evaluacija modela

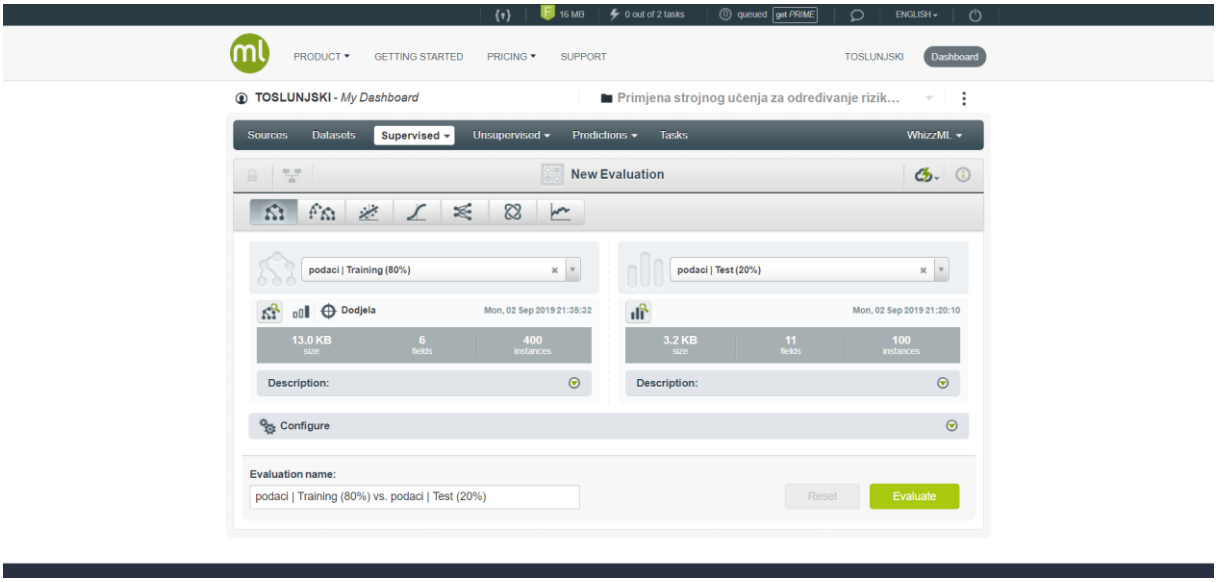
U ovom poglavlju evaluirat ću prediktivnu preciznost kreiranog prediktivnog modela. U tu svrhu koristit ću testni skup podataka koji, podsjećam, koji nije korišten u izradi modela tako da su ti podaci novi i do sad ne viđeni za model te time pogodni za njegovo testiranje.

Evaluacija se izvodi na način da se otvori kreirani model koji se nalazi pod „nadzirani“ (eng. supervised) karticom te u padajućem izborniku opcija „modeli“ (eng. models).



Slika 13. Popis modela u aplikaciji BigML

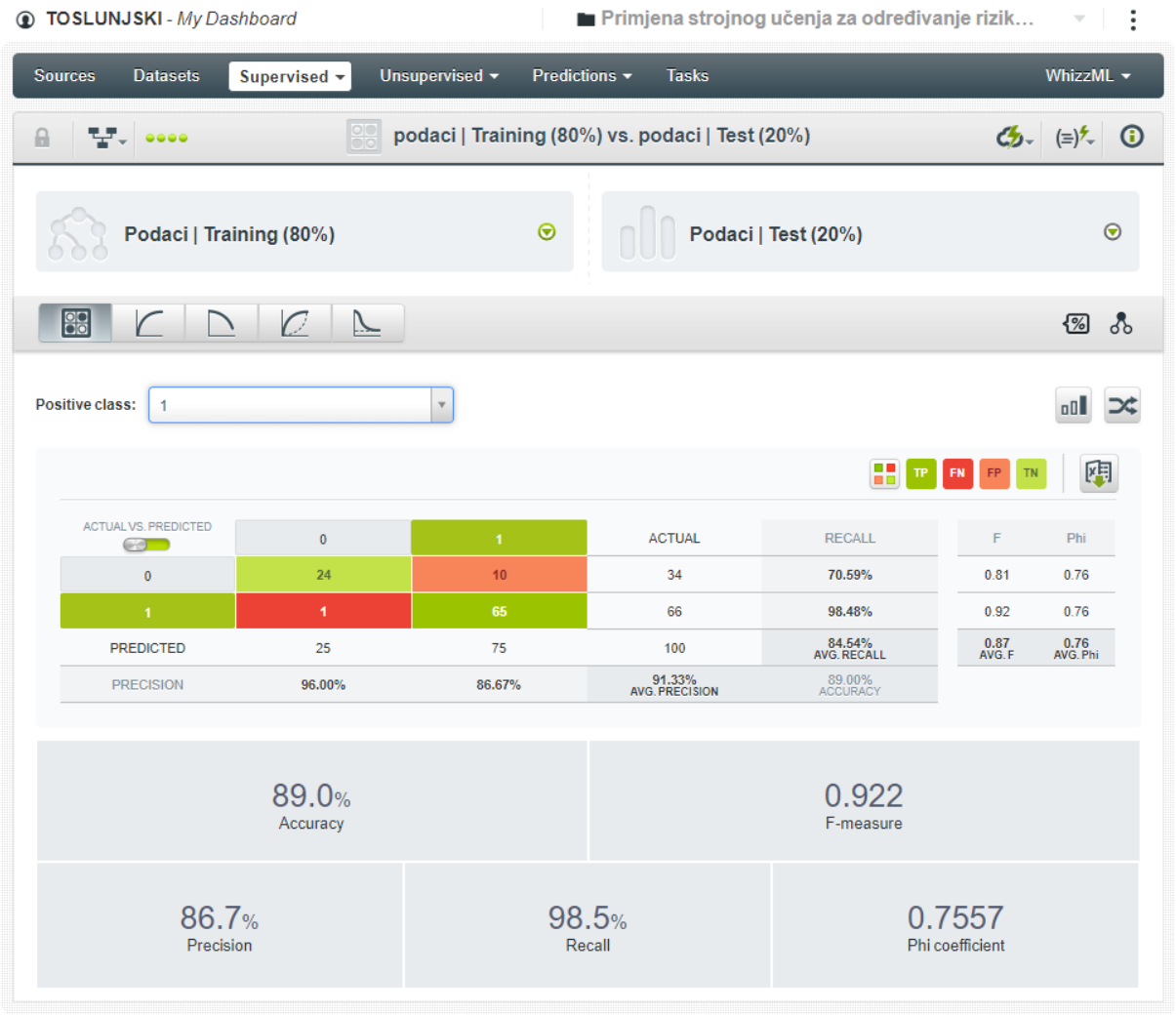
U već dobro poznatoj ikoni oblaka nalazimo opciju evaluacije. Evaluacija automatski odabire testni skup podataka, koji sam prethodno kreirao, koji sadrži 20% početnog skupa podataka koje model do sada nije vidio.



Slika 14. Evaluacijski izbornik u alatu BigML

U evaluacijskom izborniku moguća su dodatna ugađanja za kojima u ovom slučaju nema potrebe jer je već automatika alata odradila dobar posao. Konačna potvrdu i pokretanje evaluacije vrši se pritiskom na gumb „Evaluacija“.

5.2.3.1. Generalni rezultati evaluacije



Slika 15. Rezultati evaluacije modela u alatu BigML

Kao pozitivna klasa odabrana je vrijednost 1 (varijable „Dodjela) koja označava da je kandidat za kreditnu karticu zadovoljio uvjete za dodjelu kartice. Prag vjerojatnosti ostavljen je na zadanih 50%, dok je maksimalni phi koeficijent 0,9366 na koji kada bi postavili prag bi dobili agresivniji i puno precizniji model, no radi analize performansi modela na „nižem pragu“ ostavio sam prag na zadanoj vrijednosti. Na slici pod rednim brojem 15 nalaze se generalni rezultati evaluacije modela koje ću analizirati po redu:

- **Klasifikacijska tablica determiniranog modela(matrica grešaka):**
  Česta metoda za analizu prediktivnog učinka modela je matrica grešaka. To je tablica koja sadrži predviđanja i stvarne vrijednosti za ciljane klase polja, tako da mogu vizualizirati ispravne odluke kao i pogreške koje je učinio klasifikator.

Tablica 3. Matrica grešaka za klasifikacijski problem s dvije klase

Matrica greške		Primjeri	
		Negativni	Pozitivni
Prognoza	Negativna	Stvarno negativni (TN)	Lažno negativni (FN)
	Pozitivna	Lažno pozitivni (FP)	Stvarno pozitivni (TP)

Sve mjere u daljnjem tekstu se računaju na temelju podataka iz matrice grešaka koja je vidljiva na slici 15.

Vrijednosti varijabla:

Stvarno negativni(TN): 24

Stvarno pozitivni(TP): 65

Lažno negativni(FN): 10

Lažno pozitivni(FP): 1

Dakle, model je prepoznao 24 osobe koje nisu trebale dobiti kreditnu karticu i 65 osoba koje su trebale dobiti. Dok je jednu osobu koja nije trebala dobiti karticu krivo procjenio i rekao da bi trebala dobiti i za 10 osoba je rekao da ne bi trebali dobiti karticu, a radi se o kandidatima koji imaju sve uvjete za karticu i trebala bi im biti dodijeljena.

- **F – mjera (eng. F – measure):**

  F-mjera je uravnotežena harmonijska srednja vrijednost između preciznosti i odziva. F-mjera je često korisnija metrika nego točnost, budući da će slaba izvedba u bilo kojem od njih preciznosti ili opozivu rezultirat niskom vrijednosti F-mjere. Može se kretati između 0 i 1. Veće vrijednosti ukazuju na bolje performanse. Formula za F – mjeru glasi:

$$F - mjera = \frac{2 \times preciznost \times odziv}{preciznost + odziv}$$

  Na slici 15 je vidljivo da F – mjera za model iznosi 0,81 za negativne primjere i 0,92 za pozitivne primjere što daje prosječnu vrijednost F – mjere 0,87 što je dosta blizu ciljanoj vrijednosti 1.

- **Phi koeficijent (eng. Phi coefficient):**

  Koeficijent phi, koji se naziva i Mathews-ov koeficijent korelacije, koeficijent je korelacije između predviđene i stvarne vrijednosti. Vraća vrijednost između -1 i 1. Koeficijent -1 negativan povezanost predviđanja i stvarnih vrijednosti; a 0 označava da predviđanje nije bolje od nasumičnog, a koeficijent 1 ukazuje na savršeno predviđanje. Definiran je formulom:

$$Phi\ koef = \frac{T_P \times T_N - F_P \times F_N}{\sqrt{(T_P + F_P) \times (T_P + F_N) \times (T_N + F_P) \times (T_N + F_N)}}$$

Prema rezultatima sa slike 15 koeficijent je 0,76 što je prema gore navedenom dosta dobar rezultat.

- **Makro prosjeci (eng. Macro – averages):**

  Kao što je objašnjeno u prethodnim podnaslovima, mjere klasifikacije izračunavaju se za svaku klasu, osim u slučaju točnosti koja je jedina mjera koja se uvijek izračunava za cjelokupni model. BigML izračunava prosjek mjera po klasama za mjerenje ukupnih performansi modela. Te globalne statistike nazivaju se makro-prosjeci mjera jer se računaju davanjem jednakih težina svih klasa. Makro mjera za preciznost iznosi 91,33%, za točnost 89%, za F – mjeru 0,87, a za Phi koeficijent 0,76. Neke makro mijere su već komentirane u prethodnim podnaslovima, no kada se njihove vrijednosti uzmu u obzir daju još jednu pozitivnu sliku modela.

- **Točnost (eng. accuracy):**

  Točnost se izračunava kao broj ispravno razvrstanih instanci u odnosu na ukupan broj ocijenjenih slučajeva. Prema formuli:

$$Točnost = \frac{T_{POZ} + T_{NEG}}{Ukupan\ broj\ instanci}$$

  Točnost je i dalje popularna mjera za performanse modela, jer ju je vrlo lako izračunati, ali za mnogo stvarnih životnih problema previše je jednostavno i može dovesti u zabludu. Jedno od najočitijih je kada se model mora nositi s neuravnoteženim klasama. Na primjer, pretpostavimo da smo dobili 90% točnost za binarni klasifikacijski model za koji imate 900 primjeraka za jednu od klasa i 100 za drugu. Točnost od 90% je dostupna samo klasificiranjem svih 1.000 primjeraka. Točnost ima dvije mane, a to su da zanemaruje razlike između tipova grešaka i da je zavisna o distribuciji klasa u skupu podataka, a ne o karakteristikama primjera. To je razlog zašto je vrlo važno uzeti u obzir još dvije mjere, a to su preciznost i odziv.

  Točnost modela je 89% (slika 15) što je dosta dobar rezultat, dakle model će sa 89% točnošću prepoznati kandidate koji zadovoljavaju uvjete za dobivanje kreditne kartice, no više ćemo vidjeti u matrici grešaka.

- **Preciznost (eng. precision):**
  Preciznost je postotak ispravno predviđenih slučajeva u odnosu na ukupan broj slučajeva predviđenih za pozitivnu klasu. Definirana je formulama:

$$Preciznost(poz\ primjeri) = \frac{T_P}{T_P + F_P}$$

$$Preciznost(neg\ primjeri) = \frac{T_N}{T_N + F_N}$$

  Preciznost modela u negativnim primjerima iznosi 96% što je jako dobar rezultat. Dok u pozitivnim primjerima iznosi 86,67% što je prihvatljivo (slika 15). Kada usporedimo rezultate vidimo da će model u većem postotku prepoznati klijente koji nisu dobar kandidat za dodjelu kartica, samim time će se prepoznati kreditni rizik i prevenirati mogući gubitci.
- **Odziv (eng. recall):**
  Odziv je postotak ispravno klasificiranih slučajeva u odnosu na ukupne stvarne slučajeve za pozitivne klase. Definiran je formulama:

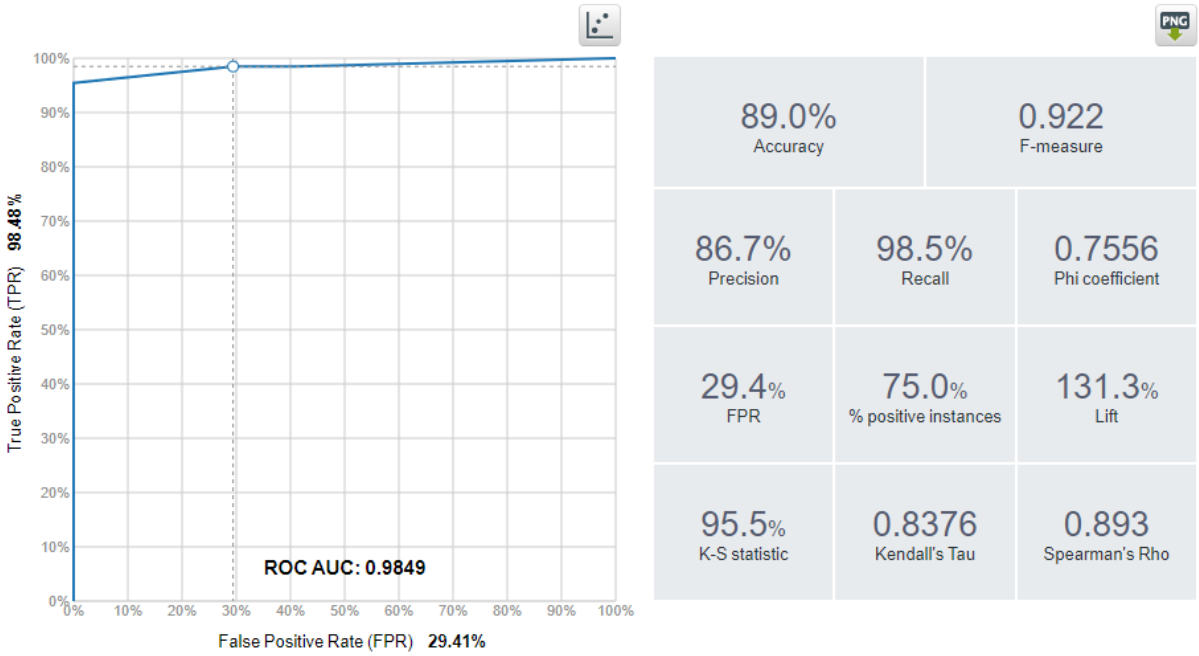
$$Odziv(poz\ primjeri) = \frac{T_P}{T_P + F_N}$$

$$Odziv(neg\ primjeri) = \frac{T_N}{T_P + F_N}$$

  Odziv za negativne primjere je 70,59% dok je za pozitivne primjere 98,49% (slika 15).

5.2.3.2. ROC krivulja

ROC (eng. Receiver Operating Characteristic) krivulja je grafički prikaz koji daje izvješće o točnosti predikcije. Prikazuje odnos opoziva (ili osjetljivosti) i specifičnosti za klasifikacijske probleme [10].

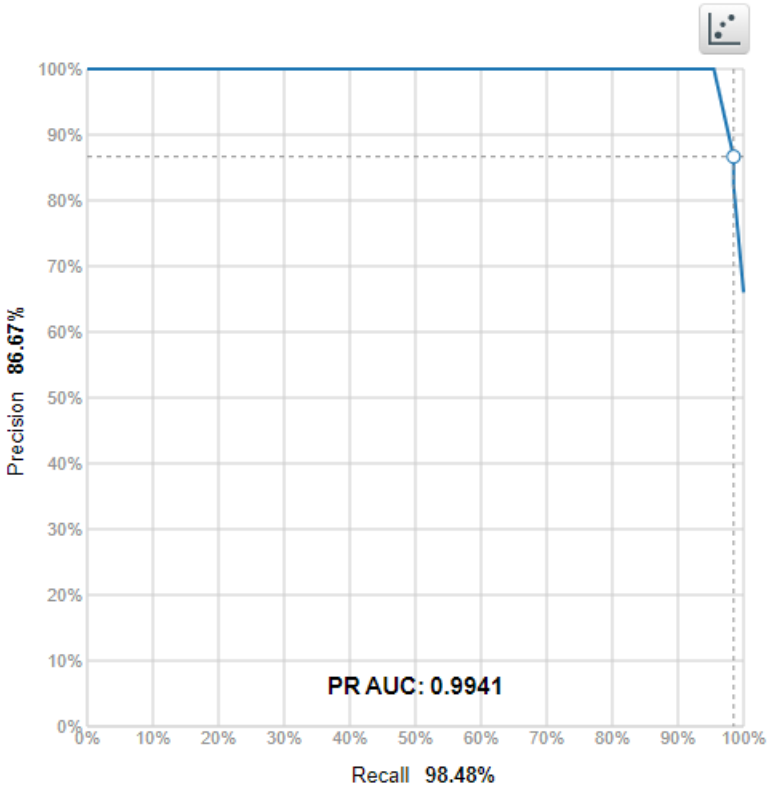


Slika 16. ROC krivulja modela u aplikaciji BigML

Na slici 16 prikazana je ROC krivulja modela. Prikazani koordinatni sustav možemo podijeliti zamišljenom dijagonalom. Sve iznad dijagonale smatra se dobrim rezultatom (gdje je TPR>FPR). Krivulja modela je cijelo vrijeme iznad zamišljene dijagonale tako da je to još jedan pokazatelj klasifikacijske moći modela. Također, ROC AUC(vrijednost jedan označava savršeni model, ali može značiti i da model ima problem prevelike glomaznosti) indikator pokazuje da se radi o modelu koji ima odlične klasifikacijske sposobnosti.

5.2.3.3. Krivulja preciznosti odziva

Krivulja preciznosti-opoziva vizualno predstavlja odnos između dvije navedene mjere koje su obje mjere za pozitivnu klasu. Preciznost i opoziv su obrnuto povezani, tj. za isti model može se povećati opoziv koristeći niži prag za pozitivnu klasu, ali obično rezultira smanjenjem preciznosti i obrnuto [10].



Slika 17. Krivulja preciznosti i opoziva modela iz aplikacije BigML

Visoka preciznost i visok opoziv predstavljeni su točkama u gornjem desnom kutu grafikona (1,1) na slici 17. Što je veća površina ispod krivulje to bolje.

5.2.3.4. Krivulja informacijskog dobitka

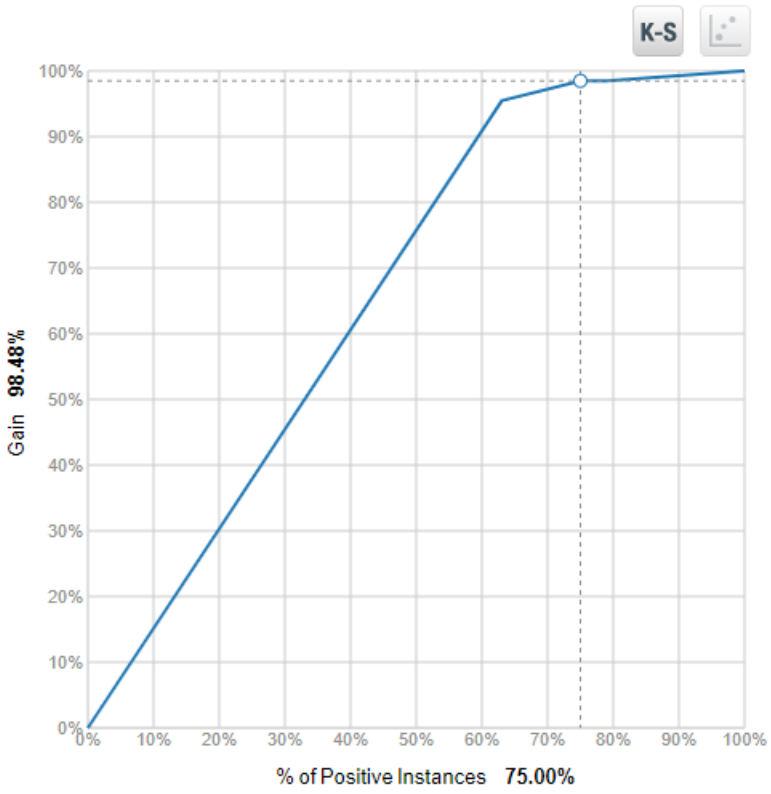
Krivulja dobitka predstavlja odnos između postotka točnih predviđanja za pozitivnu klasu i napor potreban za njihovo postizanje izmjeren kao postotak predviđenih instanci. Y os predstavlja opoziv kao i TPR (eng. true positive rate – hrv. stopa stvarno pozitivnih), a x os predstavlja postotak instanci pozitivne klase. Formule su:

$$Dobitak = Opoziv = TPR = \frac{T_P}{T_P + F_N}$$

$$\%\ Pozitivnih\ instanci = \frac{T_P + F_P}{T_P + F_P + T_N + F_N}$$

Slično krivulji ROC-a, dijagonala grafikona predstavlja rezultate slučajnog modela. Sve točke iznad dijagonale mogu se smatrati dobrim rezultatima. Što je bliža točka gornjem lijevom kut (0,1), to bolje. [10]

Na slici 18 je vidljivo da je graf cijelom dužinom iznad zamišljene dijagonale.



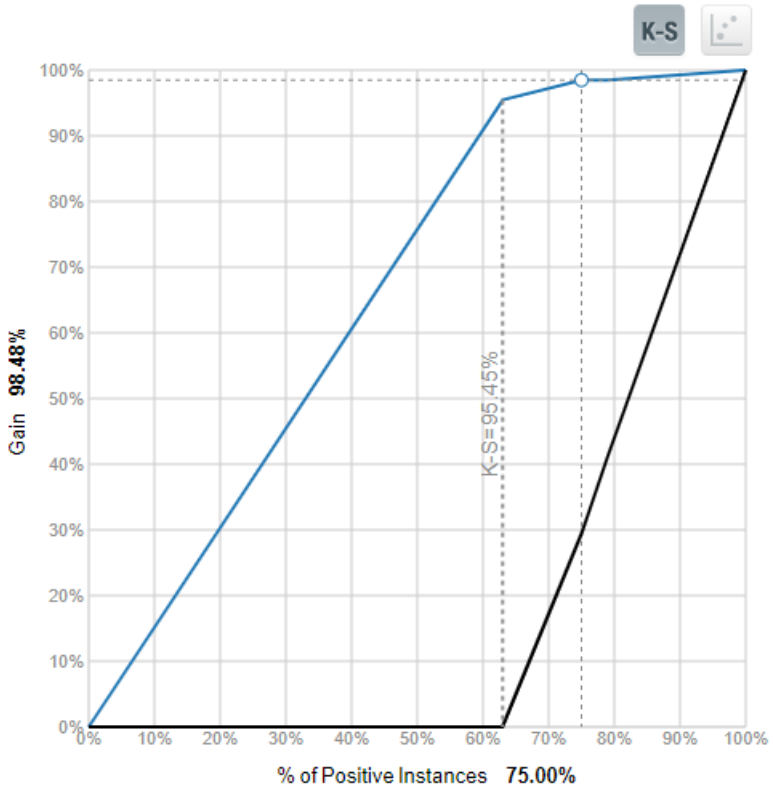
Slika 18. Krivulja dobitka modela iz aplikacije BigML

5.2.3.5. Kolmogorov-Smirnov statistika

Kratkog naziva K-S statistika, mjeri razliku između vrijednosti TPR i FPR preko svih mogućih pragova. Formula glasi:

$$K - S\ statistika = \max\left(TPR - FPR\right)$$

Ova statistika pokazatelj je koliko dobro model odvaja pozitivnu od negativne klase. Vrijednost 100% ukazuje na savršeno odvajanje i model koji sve ispravo klasificira. Što je vrijednost veća, model je kvalitetniji.



Slika 19. Graf K-S statistike modela iz aplikacije BigML

Na slici 19 vidljiv je rezultat K-S statistike od 95,45% što je iznimno dobar rezultat i ukazuje na kvalitetu modela.

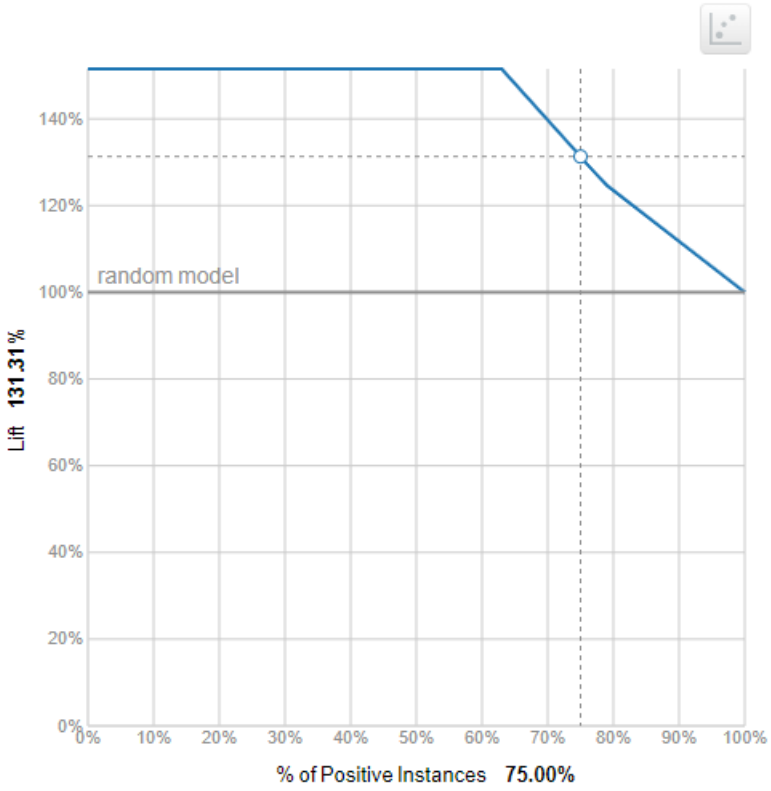
5.2.3.6. Krivulja podizanja

Krivulja podizanja pokazuje ispravnost uklapanja modela u usporedbi sa slučajnim rasporedom klasa sa danim uzorkom pozitivnih instanci. Podizanje je prikazano na y-osi i izračunava se kao omjer između rezultata koje je predvidio model i rezultata bez korištenja modela. Os x ponovo predstavlja postotak točnih predviđanja [10]. Formule za ove metrike su:

$$Podizanje = \frac{Preciznost}{\frac{Pozitivne\ instance}{Sve\ instance}} = \frac{\frac{T_P}{T_P + F_P}}{\frac{T_P + F_N}{T_P + F_P + T_N + F_N}}$$

$$\%\ Pozitivnih\ instanci = \frac{T_P + F_p}{T_P + F_P + T_N + F_N}$$

Vodoravna crta koja na grafikonu pokazuje 100% dizanje predstavlja model koji vrši slučajna predviđanja.



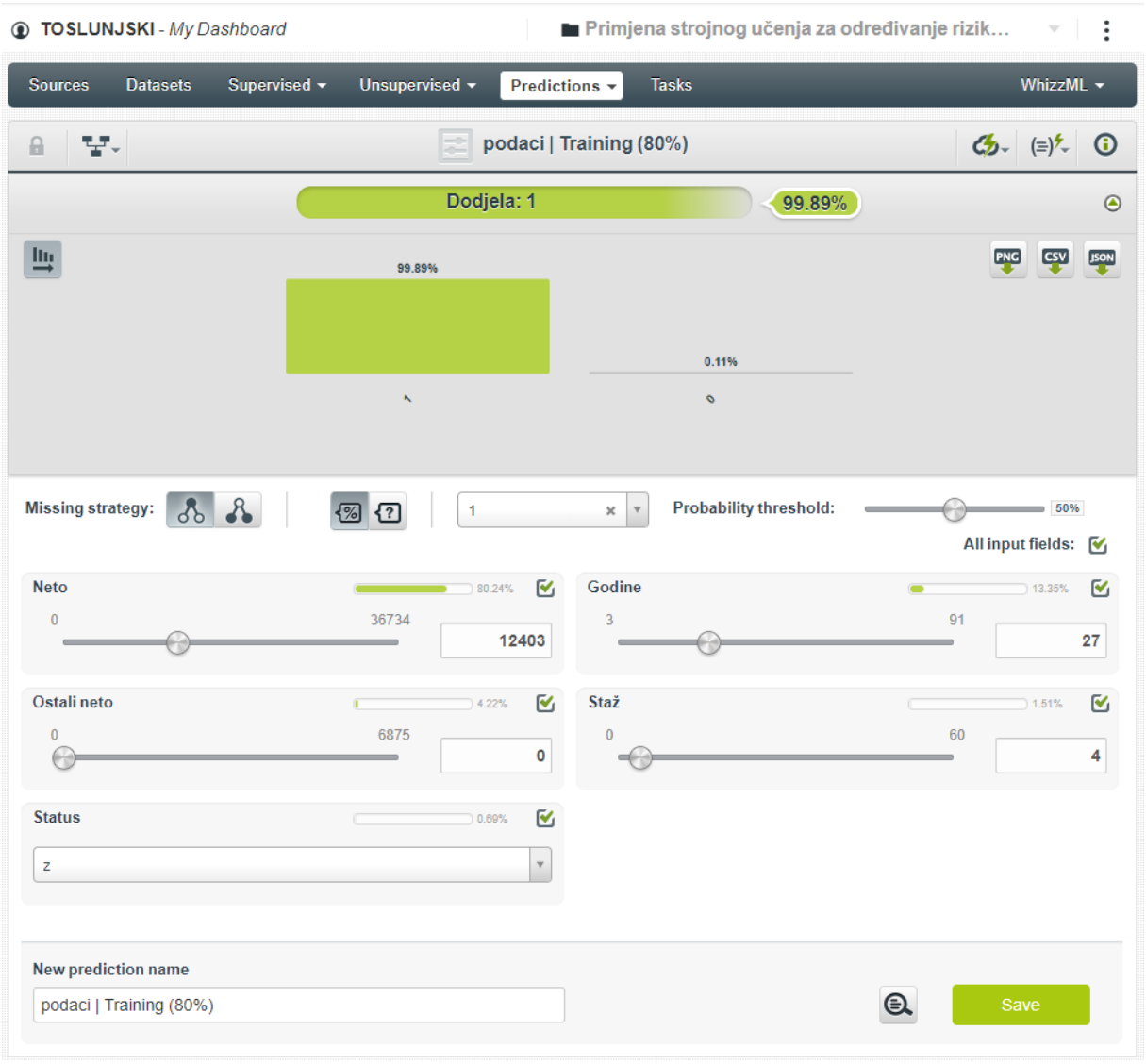
Slika 20. Krivulja podizanja modela iz aplikacije BigML

Slika 20 prikazuje krivulju podizanja modela, na njoj se vidi da model daje veliku dodanu vrijednost jer u idealnom slučaju želimo da se krivulja podizanja proširi što je više moguće u gornji lijevi kut, što ukazuje da imamo veliko dizanje.

5.2.4. Rad i primjena modela

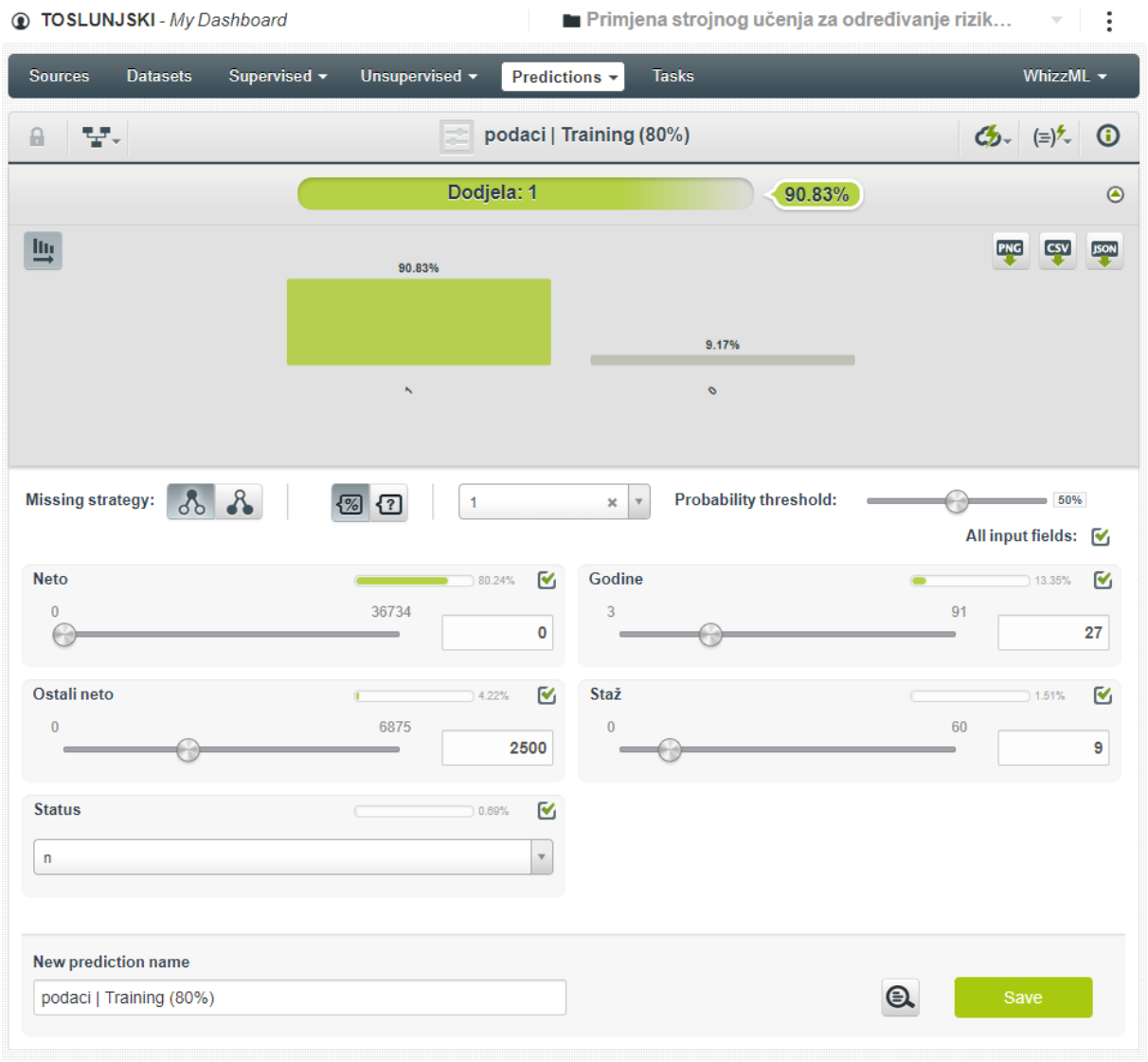
Nakon evaluacije modela isprobat ću njegov rad unosom podataka sa slike 3. Model će napraviti predikciju na temelju unesenih podataka. Predikcija se izvodi tako da se na kartici nadzirani u padajućem izborniku odabere opcija modeli. Prikazuje se popis modela te se prolaskom kursora preko popisa modela pokraj imena modela pojavljuje gumb za padajući izbornik u kojem se nalaze opcije predikcije na temelju pitanja (eng. predict by question) gdje klijent odgovara na slijedna pitanja vezana za vrijednosti pojedine varijable i opcija predikcije gdje se odjednom unesu svi podaci i pokreće model te opcija uvoza više zapisa u obliku dokumenta.

Prikazat ću drugu opciju sa već navedenim podacima.



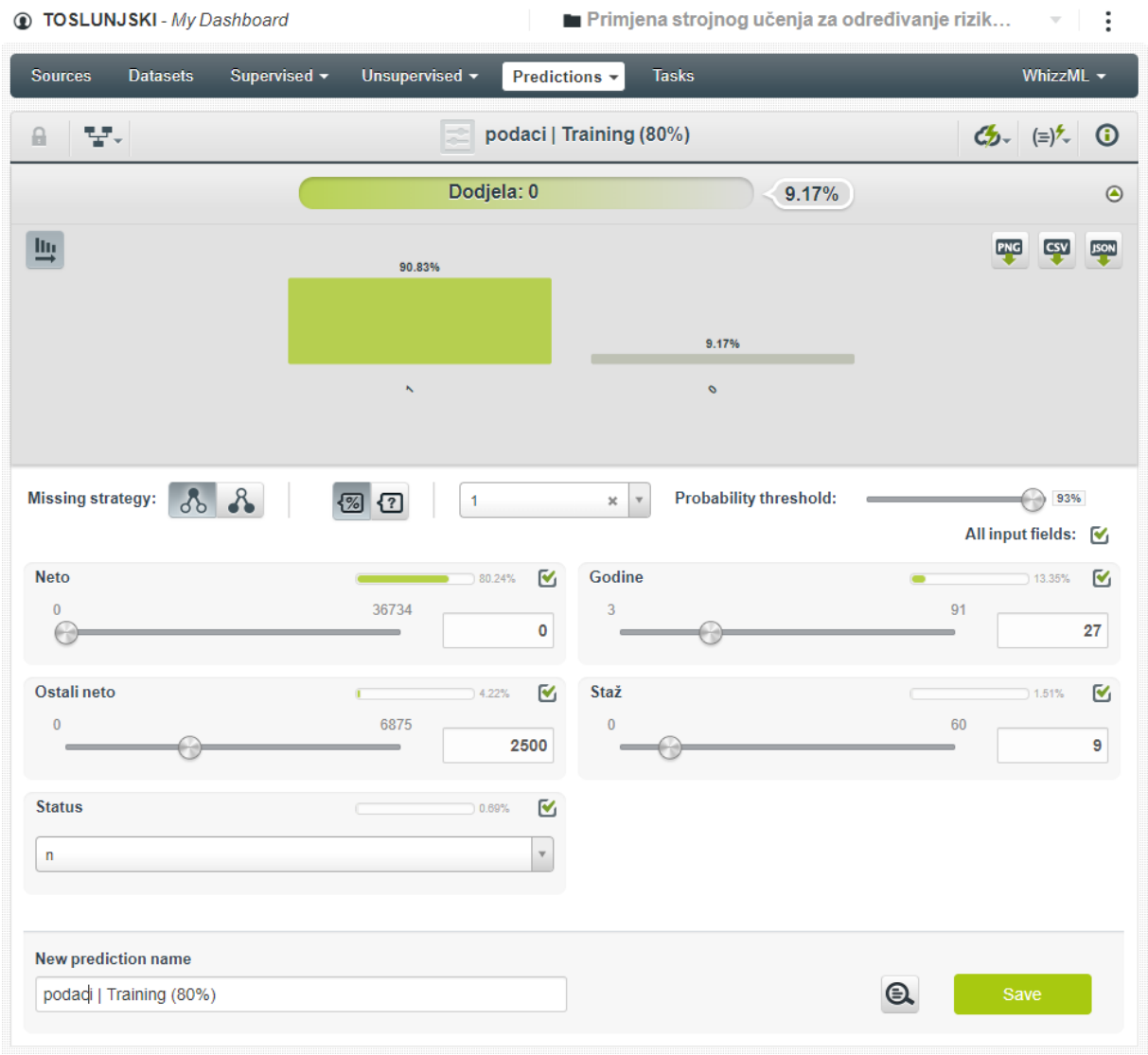
Slika 21. Prikaz pozitivne predikcije na temelju unesenih podataka u aplikaciji BigML

Na temelju unesenih podataka model je predvidio sa 99,89% vjerojatnošću da se kandidatu može dodijeliti kreditna kartica. Slijedi primjer drugog kandidata. Ovdje je situacija dosta zanimljiva jer podaci ukazuju da takvom kandidatu ne treba dodijeliti karticu, no model sa 90,83% vjerojatnošću predviđa da bi kandidat trebao dobiti kreditnu karticu. Razlog tome je što je negativna klasa dosta rijetka i neće biti često predviđena.



Slika 22. Prikaz lažne pozitivne predikcije na temelju unesenih podataka u aplikacije BigML

Tu nastupa prag vjerojatnosti (eng. probability treshold) čijim se uglađivanjem određuje prag vjerojatnosti ispod kojeg se označavaju kandidati koji nisu zadovoljili uvjete za dobivanje kreditne kartice, a iznad njega kandidati koji su zadovoljili uvjete za dobivanje kartice. Nisam nigdje naišao kako točno odrediti taj prag, no uvidio sam da ima povezanosti sa maksimalnim phi koeficijentom koji u mom modelu iznosi 0.9366. Postavljanjem praga vjerojatnosti na 93,66% u ovom slučaju sam dobio predikciju da zahtjev klijenta za kreditnu karticu treba odbiti.



Slika 23. Prikaz negativne predikcije modela nakon promjene praga vjerojatnosti u aplikaciji BigML

Dakle, uz određivanje pravog praga vjerojatnosti dobivamo vrlo moćan alat za predikcije. Kao što je navedeno i u obrađenim člancima, modeli stabla odlučivanja pokazali su zavidne rezultate tako da predstavljaju vrlo pouzdan alat za svakodnevno korištenje.

Ovakav alat bi bio od velike koristi za analitičare koji rade na procjeni zahtjeva za dodjelu kreditnih kartica. Analitičari bi imali jasno definirana pravila, ne bi trebali imati pretjerano iskustvo u procjeni zahtjeva te bi osjetili velike uštede na vremenu. Banke bi bile upoznate sa rizikom izdavanja kreditne kartice pojedinom klijentu te na taj način kontrolirale svoju izloženost kreditnom riziku.

Idealan sustav za procjenu zahtjeva bi se sastojao od web forme za popunu podataka vezanih za zahtjev za kreditnu karticu. Web forma bi podatke upisivala u bazu podataka iz koje bi se mogli uvoziti podaci, u većem broju odjednom, u model te raditi procjena na temelju koje bi se dalje vršili potrebni poslovni procesi.

Alat BigML nudi i opciju korištenja API-a koji se može integrirati u bilo koje rješenje informacijskog sustava te postati njegov modul. Nakon integracije, testiranja i verifikacije takvog modula, analitičar bi dobio alat kojim bi u trenutku mogao procijeniti zahtjev i dodatno ubrzati proces dodjeljivanja kartice korisniku.

6. Zaključak

Kroz ovaj rad htio sam analizirati te predstaviti model stabla odlučivanja u domeni donošenja odluke o dodjeli kreditne kartice klijentu na temelju podataka ispunjenih u zahtjevu za dodjelu kartice. Uz to, model bi mogao biti od velike pomoći u procjeni kreditnog rizika jer se banke izdavanjem kreditne kartice izlažu kreditnom riziku. Model bi im pomogao u definiranju veličine tog rizika te omogućio lakše donošenje odluke da li će prihvatiti taj rizik.

Da bi izradio model morao sam odabrati i upoznati alat u kojem ću izraditi model, upoznati se sa domenom kreditnog rizika, poslovanjem banaka ili kartičnih kuća sa kreditnim karticama i teorijom vezanom za stabla odlučivanja.

Kako bi se upoznao sa eventualnim postojećim problemima i rješenjima iz tih domena, analizirao sam stručne članke koji su se bavili sličnim pitanjima. Kroz analizu članaka ustanovio sam da je metoda stabla odlučivanja dosta popularna radi svoje razumljivosti, transparentnosti, pouzdanosti te velike moći predikcije. Pokazala se kao jedna od najučinkovitijih metoda za rješavanje problema iz spomenute domene.

Nakon toga sam krenuo u upoznavanje i djelomično prisjećanje teorije vezane uz stabla odlučivanja pošto sam se s tom metodom već susretao prilikom svog akademskog puta kroz određene kolegije na fakultetu. Kroz analizu stručne literature i raznih članaka i radova na tu temu skupio sam određenu teoretsku podlogu te krenuo u potragu za alatom koji će mi omogućiti izradu modela stabla odlučivanja.

Nakon određene potrage uz prijedlog mentorice odlučio sam se za alat BigML. On se pokazao kao pun pogodak jer besplatno nudi dosta mogućnosti koje su pristupačne, razumljive i pomalo intuitivne. Alat nudi široku paletu raznih modela, njihovih analiza i evaluacija potkrepljenih jasnim i transparentnim vizualnim prikazima.

Za izradu modela bio mi je potreban skup podataka koji sam, nakon duge potrage po edukacijskim stranicama na internetu, odlučio napraviti sam simuliranjem podataka dodijeljenih određenim varijablama koje sam definirao na temelju analize formi zahtjeva određenih kartičnih kuća koje posluju u Republici Hrvatskoj. Kada sam izradio skup podataka pristupio sam izradi modela.

Model je bilo dosta lako izraditi zbog jednostavnosti i pristupačnosti alata koji je potkrepljen kvalitetnom dokumentacijom među kojom sam naišao na detaljne upute koje su mi bile od velike koristi. Nakon izrade modela, napravio sam evaluaciju modela te rezultate evaluacije prikazao u radu.

Rezultati su bili jako dobri što pokazuje da model ima iznimnu predikcijsku moć za predstavljeni problem te opravdava svoju popularnost u znanstvenim krugovima koju sam imao prilike vidjeti kroz analizu znanstvenih članaka. Model bi se mogao koristiti, uz vjerojatno potrebna dodatnog uglađivanja i testom nad većim skupom podataka, u realnim sustavima. Alat BigML nudi integraciju svojih mogućnosti kroz API u bilo koji sustav što je danas velika prednost za jednu aplikaciju takvog tipa.

Kroz izradu rada upoznao sam metodu stabla odlučivanja, upoznao se sa radom u aplikaciji BigML te domenom kreditnog skoringa, rizika te kartičnog poslovanja. Metoda stabla odlučivanja primjenjiva je u raznim domenama stoga mi je drago da sam stekao određena znanja o njoj koja ću, nadam se, imati prilike negdje i primijeniti.

Popis literature

[1] L. Kvesić, »Primjena stabla odlučivanja u kreditnom skoringu,« *Ekonomski vjesnik,* pp. 382-390, 2013.

[2] N. Šarlija, »Efos Unios,« 2008. [Mrežno]. Available: http://www.efos.unios.hr/kreditna-analiza/wp-content/uploads/sites/252/2013/04/4_rizici-u-bankama.doc.pdf. [Pokušaj pristupa 25. 8. 2019.].

[3] H. T. B. R. N. H. T. Sikavica P., Poslovno odlučivanje, Školska knjiga, 2014., p. 576.

[4] D. Hruška, »Slideserve,« 2014. [Mrežno]. Available: https://www.slideserve.com/hana/stablo-odlu-ivanja. [Pokušaj pristupa 29. 8. 2019.].

[5] M. Zekić-Sušac, »EFOS,« 2017. [Mrežno]. Available: http://www.efos.unios.hr/sustavi-poslovne-inteligencije/wp-content/uploads/sites/192/2017/10/P4_Stabla-odlucivanja-2017.pdf. [Pokušaj pristupa 16. 8. 2019.].

[6] Big ML team, »Big ML,« [Mrežno]. Available: https://support.bigml.com/hc/en-us/articles/206616279-What-kind-of-algorithm-does-BigML-use-to-build-decision-tree-models-and-how-does-it-work-. [Pokušaj pristupa 27. 8. 2019.].

[7] Big ML support team, »Big ML,« [Mrežno]. Available: https://support.bigml.com/hc/en-us/articles/206616279-What-kind-of-algorithm-does-BigML-use-to-build-decision-tree-models-and-how-does-it-work-. [Pokušaj pristupa 1. 9. 2019.].

[8] J. Galindo i P. Tamayo, »Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications,« *Computional Economics,* br. 15, pp. 107-143, 2000.

[9] A. E. Khandani, A. J. Kim i A. W. Lo, »Consumer credit-risk models via machine-learning algorithms,« *Journal of Banking & Finance,* br. 34, pp. 2767-2787, 2010.

[10] The BigML team, »BigML,« [Mrežno]. Available: https://static.bigml.com/pdf/BigML_Classification_and_Regression.pdf?ver=ebd8c29. [Pokušaj pristupa 15. 4. 2019.].

[11] bigmlcom, »YouTube,« 23 8 2017. [Mrežno]. Available: https://www.youtube.com/watch?v=hnt7z24wvxs. [Pokušaj pristupa 11. 7. 2019.].

[12] Nepoznat, »web math,« [Mrežno]. Available: https://web.math.pmf.unizg.hr › nastava › index.php › download_file. [Pokušaj pristupa 30. 8. 2019.].

[13] Z. Bohaček, N. Šarlija i M. Benšić, »EFOS,« 2013. [Mrežno]. Available: http://www.efos.unios.hr/nsarlija/wp-content/uploads/sites/88/2013/04/upotreba-kredit-skoring-modela_r.pdf. [Pokušaj pristupa 28. 8. 2019.].

[14] F. Hojsak, »BiB Irb,« 2017. [Mrežno]. Available: https://bib.irb.hr/datoteka/899229.Hojsak_-_Rizik_v4.pdf. [Pokušaj pristupa 20. 8. 2019.].

[15] J. Brownlee, »machinelearningmastery,« 7 4 2014. [Mrežno]. Available: https://machinelearningmastery.com/bigml-tutorial-develop-your-first-decision-tree-and-make-predictions/. [Pokušaj pristupa 3. 9. 2019.].

[16] L. Tokić, »NSK,« 9 2017. [Mrežno]. Available: https://zir.nsk.hr/islandora/object/pmf%3A4490/datastream/PDF/view. [Pokušaj pristupa 27. 8. 2019.].

[17] S. Littler, »Select statistics,« [Mrežno]. Available: https://select-statistics.co.uk/blog/cumulative-gains-and-lift-curves-measuring-the-performance-of-a-marketing-campaign/. [Pokušaj pristupa 27. 8. 2019.].

[18] T. Srivastava, »Analyticsvidhya,« 6 8 2019. [Mrežno]. Available: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/. [Pokušaj pristupa 27. 8. 2019.].

Popis slika

Slika 1. Grafički prikaz stabla odluke [4] 6

Slika 2. Klasifikacijsko stablo odlučivanja [5] 7

Slika 3. Primjer zapisa iz skupa podataka 10

Slika 4. Prikaz alata BigML 11

Slika 5. Stablo odlučivanja dobiveno CHAID algoritmom pomoću SPSS statističkog paketa [1]16

Slika 6. Glavni ekran aplikacije BigML nakon prijave u aplikaciju 24

Slika 7. Prikaz kontrolne ploče u aplikaciji BigML sa karticama 25

Slika 8. Pregled izvora podataka u aplikaciji BigML 25

Slika 9. Popis skupova podataka u aplikaciji BigML 26

Slika 10. Detalji skupa trening podataka u aplikaciji BigML 27

Slika 11. Prikaz prediktivnog modela stabla odlučivanja na temelju trening podataka u aplikaciji BigML 28

Slika 12. Detaljniji prikaz čvorova grane stabla odlučivanja u aplikaciji BigML 29

Slika 13. Popis modela u aplikaciji BigML 30

Slika 14. Evaluacijski izbornik u alatu BigML 31

Slika 15. Rezultati evaluacije modela u alatu BigML 32

Slika 16. ROC krivulja modela u aplikaciji BigML 36

Slika 17. Krivulja preciznosti i opoziva modela iz aplikacije BigML 37

Slika 18. Krivulja dobitka modela iz aplikacije BigML 38

Slika 19. Graf K-S statistike modela iz aplikacije BigML 39

Slika 20. Krivulja podizanja modela iz aplikacije BigML 41

Slika 21. Prikaz pozitivne predikcije na temelju unesenih podataka u aplikaciji BigML 42

Slika 22. Prikaz lažne pozitivne predikcije na temelju unesenih podataka u aplikacije BigML 43

Slika 23. Prikaz negativne predikcije modela nakon promjene praga vjerojatnosti u aplikaciji BigML 44

Popis tablica

Tablica 1. Opis korištenih varijabli 9

Tablica 2. Usporedba preciznosti i vremenske razlike u klasifikacijsko regresijskog modela stabla[7] 19

Tablica 3. Matrica grešaka za klasifikacijski problem s dvije klase 33