Projektni prijedlog

- Predviđanje stečaja poduzeća u Hrvatskoj -

David Bojanić, Domagoj Demeterfi

04/2018

Sadržaj

1 Sažetak **2**

2 Uvod **3**

2.1 Ciljevi projekta 3

2.2 Dosadašnja istraživanja 4

3 Metodologija **6**

3.1 Stabla odluke 6

3.2 Ansambl stabala odluke 7

3.3 Extreme Gradient Boosting 8

3.4 Sintetičke značajke 8

3.5 Ansambl boosted stabala za predviđanje stečaja 9

3.6 Metode evaluacije 9

4 Eksperiment **13**

4.1 Podatkovni skup 13

4.2 Eksperimentalni setup 14

4.3 Evaluacija modela 14

4.4 Interpretacija 15

5 Literatura **16**

Predviđanje stečaja poduzeća u Hrvatskoj

1 Sažetak

Predviđanje stečaja je područje od velikog interesa u ekonomiji. Cilj projekta je postaviti prediktivni model koji koristeći ekonometrijske mjere poduzeća u Hrvatskoj predviđa ulazak tog poduzeća u postupak predstečajne nagodbe. Uz to cilj je dobiveni model interpretirati s ekonomskog stajališta. Ovaj projektni prijedlog sadrži opis problema, motivaciju njegovog rješavanja, konkretne ciljeve projekta i detaljan prijedlog rješenja.

2 Uvod

Predviđanje stečaja poduzeća od iznimne je važnosti pri donošenju ekonomskih odluka. Stanje poduzeća, bilo malog ili velikog, od interesa je za lokalnu zajednicu, sudionike u industriji i investitore, ali i za zakonodavce te globalnu ekonomiju. Stoga ne čudi što taj problem već dugo privlači pozornost istraživača.

U ovom projektu razvijamo model za predviđanje ulaska poduzeća u postupak predstečajne nagodbe. Projekt je inspiriran radom [1] u kojem su autori razvili okvir za predviđanje stečaja poduzeća temeljen na modelu ansambla boosted stabala koji je treniran tehnikom Extreme Gradient Boosting. Osnova našeg rada je također spomenuti model (detaljnije o modelu u Metodologija) no umjesto predviđanja stečaja mi pokušavamo predvidjeti ulazak poduzeća u postupak predstečajne nagodbe. U [1] autori su svoj model testirali na podacima za poljska poduzeća u sektoru proizvodnje, a mi ćemo svoj testirati na podacima o hrvatskim poduzećima u građevinskom sektoru.

Osim samog razvoja modela predviđanja ulaska u postupak predstečajne nagodbe u projektu bismo veliki naglasak stavili i na interpretaciju dobivenog modela s ekonomskog stajališta.

2.1 Ciljevi projekta

Obično se u literaturi koja se bavi predviđanjem propadanja poduzeća predviđa ulazak poduzeća u stečaj. Mi smo kroz razgovor s ekspertom iz ovog područja i zbog dostupnih podataka odlučili da bi za hrvatski slučaj bilo zanimljivo predviđati ulazak u postupak predstečajne nagodbe. Naime, zbog zakonske regulative specifične za Hrvatsku poduzeća u Hrvatskoj su zapravo niz godina prije samog ulaska u stečaj u velikim problemima.

Dakle, osnovni cilj ovog projekta je dobiti model za predviđanje ulaska poduzeća u postupak predstečajne nagodbe. Unutar modela poduzeća su opisana značajkama koje su dobivene iz njihovih bilanci iz određene godine. Više o značajkama nalazi se u dijelu Podatkovni skup. Želimo dobiti model koji će na temelju tih značajki predviđati hoće li poduzeće u nekom budućem trenutku ući u postupak predstečajne nagodbe. Potrebno je specificirati u kojem periodu predviđamo. Po uzoru na [1] zapravo ćemo rješavati pet zadataka binarne klasifikacije u ovisnosti o periodu predviđanja: poduzeće je ušlo u postupak predstečajne nagodbe nakon jedne, dvije, tri, četiri ili pet godina. Potencijalne mjere uspješnosti ovog dijela projekta objašnjene su u poglavlju Metode evaluacije dok je konkretan način njihove primjene opisan

u poglavlju Evaluacija modela.

Kao što smo već spomenuli osim same prilagodbe modela u ovom projektu bavit ćemo se i njegovom interpretacijom. Detalji o načinu interpretacije nalaze se u poglavlju Interpretacija. Također se nadamo da ćemo imati priliku o rezultatima iz ovog segmenta projekta prodiskutirati s ekspertom iz ovog područja.

Zanimljivo bi u daljnjem istraživanju bilo umjesto predviđanja samo binarnog ishoda hoće li poduzeće ući u postupak predstečajne nagodbe predviđati i trajanje predstečajne nagodbe i/ili iznos tražbine nakon prijeboja u predstečajnoj nagodbi. Također proširenje istraživanja je moguće u vidu dodavanja novih značajki koje će opisivati makroekonomske uvjete, izloženost poduzeća drugim poduzećima i slično.

2.2 Dosadašnja istraživanja

Prvi pokušaji formalnog predviđanja stečaja javljaju se početkom 20. stoljeća kada su predloženi prvi ekonometrijski indikatori za opis mogućnosti predviđanja propasti poduzeća. Šezdesete godine predstavljaju prekretnicu u istraživanju ranog otkrivanja uzroka propasti poduzeća - počinju se primjenjivati statistički modeli u svrhu predviđanja stečaja (veliki naglasak stavljen je na generalizirane linearne modele). U doba kada su velike količine podataka postale dostupne ispostavilo se da tradicionalno korišteni linearni modeli ne mogu odraziti netrivijalne veze među ekonomskim pokazateljima. Od devedesetih godina 20. stoljeća umjetna inteligencija i strojno učenje postali su značajan smjer istraživanja predviđanja stečaja. Među najpopularniji pristupima su metoda potpornih vektora, neuronske mreže te u zadnje vrijeme ansambli klasifikatora. U nastavku slijedi par izdvojenih radova koji se bave sličnom problematikom kao i mi.

Zieba et al. su u [1] vrlo uspješno primijenili ansambl boosted stabala u svrhu predviđanja stečaja poljskih poduzeća. Svoj model su eksperimentom usporedili s pristupima uobičajenim u području predviđanja stečaja te dobili značajno bolje rezultate. Kao opis poduzeća oni koriste 64 financijska omjera korištena u integriranim financijskim modelima i financijskoj analizi, a izračunati su iz podataka koji se nalaze u financijskim izvještajima poduzeća (popis značajki se može naći u Table 2 u [1]). Oni u svom radu predlažu korištenje dodatnih sintetičkih značajki u svrhu poboljšanja predikcijskih svojstava ansambla stabala. Sintetičke značajke računaju se u svakom koraku boostinga kombinirajući postojeće značajke nekom od osnovnih aritmetičkih operacija $(+, -, *, /)$ (detalji o sintetičkim značajkama su u Sin-

tetičke značajke). Evaluacijom važnosti značajki navode kao korisne u predviđanju stečaja poduzeća prilagođen udio kapitala u financiranju imovine, koeficijent tekuće likvidnosti i koeficijent obrtaja obveza. Također zaključuju da bi korisni mogli biti i koeficijenti profitabilnosti, koeficijent financijske poluge te neki drugi. Analizom važnosti sintetičkih značajki zaključuju da su važne operativne performanse, profitabilnost firme te financijska poluga.

Alfaro et al. u [2] uspoređuju uspješnost primjene AdaBoost strategije kombiniranja stabala u ansambl i neuronske mreže za problem predviđanja stečaja. Njihov eksperiment pokazao je superiornost ansambla stabala nad individualnom neuronskom mrežom. Kao prediktivne varijable koriste 13 uobičajenih financijskih omjera temeljene na računovodstvenim podacima, ali i neke kvalitativne varijable kao što su veličina firme, sektor i vlasnička struktura (za razliku od češće prakse gdje se poduzeća podijele u grupe po npr. sektoru kao u [1] oni sektor u kojem poduzeće posluje uzimaju kao prediktivnu varijablu). Točan popis tih varijabli može se naći u Table 2 u [2]. U okviru neuronske mreže analizirali su relativni doprinos svake input varijable globalnoj performansi pomoću analize senzitivnosti. Prema toj analizi došli su do zaključka da su najvažnije značajke redom po važnosti efikasnost (prodaja kroz ukupna imovina), sektor, varijabla koja procjenjuje veličinu firme, vlasnička struktura te razina zaduženosti. Analizom modela dobivenog AdaBoost strategijom zaključuju da su najvažnije značajke redom po važnosti ekonomska profitabilnost (pokazuje uspješnost firme u korištenju imovine), razina zaduženosti, efikasnost i varijabla koja procjenjuje veličinu firme. Zadnje četiri značajke pojavljuju se i kod neuronske mreže no u drugom redoslijedu po važnosti.

Rujoub et al. u [3] ocjenjuju korisnost varijabli vezanih uz novčani tok u predviđanju stečaja. Multrivarijatnom diskriminantnom analizom isputuju 3 hipoteze: diskriminantna sposobnost podataka o novčanom toku (u obliku financijskih omjera) za predviđanje stečaja statistički je značajna, točnost klasifikacije temeljene na informacijama o novčanom toku veća je od klasifikacijske točnosti temeljene na konvencionalno korištenim računovodstvenim podacima (kao što su koeficijenti profitabilnosti, zaduženosti i slično) te korištenje podataka o novčanom toku zajedno s tradicionalnim podacima iz računovodstva može poboljšati točnost klasifikacije za predviđanje stečaja. U testiranju tih hipoteza došli su do zaključaka da korištenje podataka o novčanom toku bolje predviđa stečaj nego tradicionalno korišteni računovodstveni podaci te da korištenje podataka o novčanom toku zajedno s tradicionalnim računovodstvenim podacima poboljšava moć predviđanja računovodstvenih podataka korištenih u prethodnim istraživanjima.

Liang et al. u [4] ispituju diskriminatornu moć dobivenu kombiniranjem različitih kategorija financijskih koeficijenata (FR) i indikatora korporativnog upravljanja (CGI). Preciznije, uzeli su u obzir sedam kategorija FR-a (koeficijenti profitabilnosti, kapitalne strukture, obrtaja, novčanog toka te rasta) i pet kategorija CGI-a (struktura upravnog odbora, vlasnička struktura, prava od novčanih tokova te zadržavanje ključnog osoblja). Da bi odredili najbolju kombinaciju FR-a i CGI-a koristili su podatke o poduzećima u Tajvanu. Njihovi rezultati pokazuju da kombinacija FR-a i CGI-a može poboljšati performanse modela kada se usporede s modelom koji koristi samo FR-ove. Nadalje, zaključuju da su najvažnije značajke za predviđanje stečaja FR kategorije solventnost i profitabilnost te CGI kategorije struktura upravnog odbora i upravljačka struktura. No, korisnost CGI značajki ovisi o tržištu. Proveli su analognu analizu za kinesko tržište te došli do zaključka da prediktivne performanse kombinacije FR-a i CGI-a na tom tržištu nisu bolje od korištenja samo FR-a. Oni kao razloge tome navode drugačije definicije stečaja na tim tržištima te da je opseg u kojemu su CGI-ovi povezani s karakteristikama kompanije ovisan o tržištu.

3 Metodologija

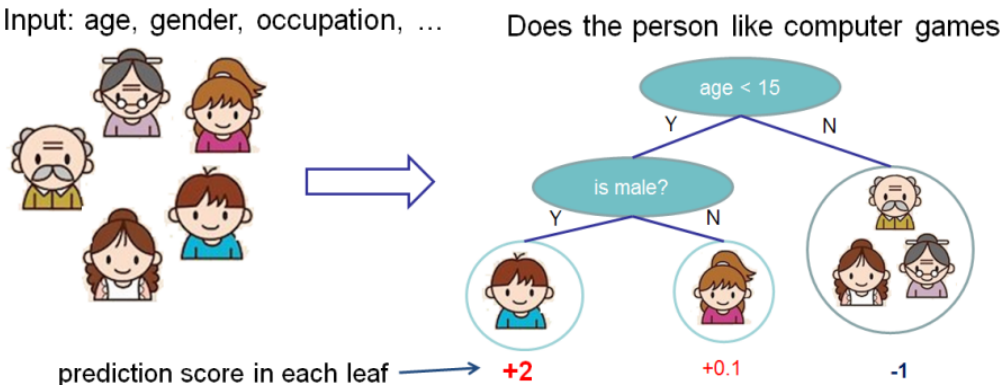
Naš rad temelji se na modelu ansambl boosted stabala treniran tehnikom extreme gradient boostinga s generacijom sintetičkih značajki (EXGB model) prezentiranom u [1]. U ovom poglavlju pokušati ćemo ukratko opisati model, tj. njegove komponente s ciljem razumijevanja kako i zašto model radi, na intuitivnoj razini.

3.1 Stabla odluke

Stablo odluke je alat koji pomaže pri donošenju odluka, tj. modelira proces donošenja odluka i njihovih mogućih posljedica. Učenje stablom odluke koristi stabla odluke kao prediktivni model da bi od opažanja o nekom predmetu (prikazanih granama stabla) došli do zaključaka o ciljanoj vrijednosti predmeta od interesa (prikazanim u listovima stabla). To je jedan pristup modeliranja korišten u statistici, rudarenju podataka i strojnom učenju.

Na slici 1 je primjer stabla odluke koje na temelju nekih značajki osobe (starost, spol, zanimanje...) klasificira voli li osoba video igre ili ne.

Klasificiramo osobe u različite listove stabla te im pridružujemo neku težinu na odgovarajućem listu. Na slici je zapravo primjer klasifikacijskog i regresijskog stabla (CART). Za



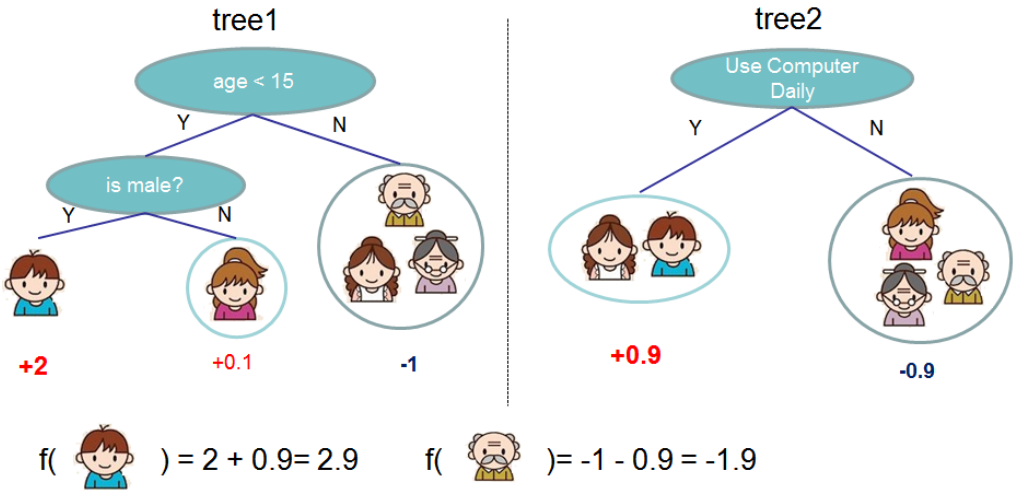
Slika 1: Slika preuzeta s Introduction to Boosted Trees

razliku od običnih stabala odluke kod kojih list sadrži samo vrijednost odluke, u CART-u je listu pridružena realna težina. To nam omogućuje širu interpretaciju i tehnički olakšava boosting (pri optimizaciji).

3.2 Ansambl stabala odluke

Obično jedno stablo nije dovoljno jako da bi se koristilo u praksi. Zapravo se koristi takozvani model ansambla stabala koji je skup jednostavnih stabala odluke. Ansambl stabala sumira predviđene težine na različitim stablima.

Ideja učenja ansamblom je istrenirati i kombinirati obično slabe klasifikatore kako bi dobili bolje prediktivne performanse.



Slika 2: Slika preuzeta s Introduction to Boosted Trees

Na slici 2 je primjer ansambla s dva stabla. Predviđene težine svakog stabla zbrojene su kako bi se dobio konačan rezultat. Važno je uočiti ideju da se stabla u ansamblu komplementiraju.

3.3 Extreme Gradient Boosting

Nakon što smo uveli model postavlja se pitanje kako ga istrenirati, odnosno kako naučiti ova stabla. Odgovor je, kao uvijek kod nadziranog učenja, definirati funkciju cilja te ju optimizirati.

Gradient boosting je tehnika učenja (treniranja) ansambla stabala koja stvara predikcijski model "korak po korak" (u svakom koraku algoritma već naučenom skupu stabala dodaje se novo stablo takvo da se cjelokupni model poboljša) kao i druge boosting metode. Gradient boosting generalizira druge boosting metode jer dopušta optimizaciju proizvoljne diferencijabilne funkcije cilja.

Extreme gradient boosting koristi formalizaciju modela s većom regularizacijom kako bi se kontrolirao over-fitting, što mu obično daje bolje performanse od običnog gradient boostinga.

3.4 Sintetičke značajke

Zieba et al. u [1] predlažu korištenje sintetičkih značajki u svrhu poboljšanja predikcijskih svojstava ansambla stabala. Sintetičke značajke računaju se u svakom koraku boostinga kombinirajući postojeće značajke nekom od osnovnih aritmetičkih operacija $(+, -, *, /)$.

Na svaku sintetičku značajku možemo gledati kao na jedan regresijski model. Drugi pogled na sintetičke značajke je da su one na neki način analogon skrivenih jedinica u neuronskim mrežama, ali način na koji se izvlače je potpuno drukčiji.

Svrha sintetičkih značajki je kombinirati ekonomske indikatore predložene od stručnjaka u kompleksne značajke. Te kompleksne značajke mogu imati bolji utjecaj na predviđanje nego tipični ekonomski faktori.

Sintetičke značajke generirane su slučajnim odabirom dvije postojeće značajke te slučajnim odabirom aritmetičke operacije koja će na njima biti izvršena. Vjerojatnost odabira neke značajke ovisi o popularnosti (broju pojavljivanja) te značajke u već postojećim stablima (što veća popularnost veća je vjerojatnost odabira te značajke). Na to možemo gledati kao na evolucijski pristup koji odabire "najsnažnije" roditelje značajke za potomak značajku. Operacija

se uniformno bira iz skupa $\{+, -, *, /\}$.

3.5 Ansambl boosted stabala za predviđanje stečaja

Procjenitelji ekonomskih indikatora koji opisuju kompanije karakterizirani su visokom varijancom uzrokovanom relativno malim uzorkom. U praksi to znači da je većina vrijednosti akumulirana u nekom uskom segmentu no postoje kompanije koje su opisane relativno visokim/niskim vrijednostima tih značajki. Stoga primjena gradijentnih modela kao što su neuronske mreže ili logistička regresija vodi do problema s treniranjem pa i lošim predviđanjem. Taj problem teško je riješiti i kada se podatci normaliziraju ili standardiziraju. Nasuprot ovim pristupima modeli temeljeni na ansamblu stabala uzimaju u obzir red vrijednosti značajki, a ne same vrijednosti. Stoga su otporni na ogromne vrijednosti ekonomskih indikatora i ne zahtijevaju preprocesiranje podataka.

Drugi značajan problem kod primjene metoda strojnog učenja na predviđanja stečaja je i fenomen neuravnoteženosti podataka - obično postoji puno više uspješnih kompanija nego onih u stečaju. EXGB model nije osjetljiv na ovaj problem.

Nedavno je pokazano da se ansambl klasifikator može uspješno primijeniti za predviđanje stečaja ([5]) te da je značajno bolji od drugih metoda ([2]).

Osim što model ima dobra predikcijska svojstva smatramo da je EXGB model zadovoljavajuć i po kriteriju razumljivosti i interpretabilnosti. Nadamo se da smo u prethodnom poglavlju uspjeli ukratko na intuitivnoj razini objasniti kako model radi. Model može poslužiti i za određivanje važnosti određenih značajki kao uzroka propadanja poduzeća. Više o interpretaciji modela može se vidjeti u Interpretacija.

3.6 Metode evaluacije

Metode evaluacije su metode kojima procjenjujemo prediktivnu moć algoritma. Među empirijskim procjenama greške algoritma postoje razne tehnike probira (en. resampling) kojima dijelimo skup podataka na manje skupove na kojima onda procjenjujemo grešku. Tim tehnikama pripadaju Train & Test metoda, unakrsna validacija (en. Cross Validation) i LOOCV (en. Leave-One-Out-Cross-Validation).

Train & test je najprihvaćenija metoda među algoritmima strojnoga učenja. Ta metoda dijeli podatkovni skup na dio na kojemu učimo algoritam, te dio na kojem ga evaluiramo računajući koliko predikcije algoritma odstupaju od stvarnih podataka.

Kod unakrsne validacije, dijelimo skup podataka na n jednakih dijelova te ga iterativno treniramo na njih $n-1$, a validiramo na preostalom skupu. Ovakva metoda je bolja od procjene samih reziduala algoritma jer iz reziduala ne možemo procijeniti koliko dobro će algoritam na novim podacima.

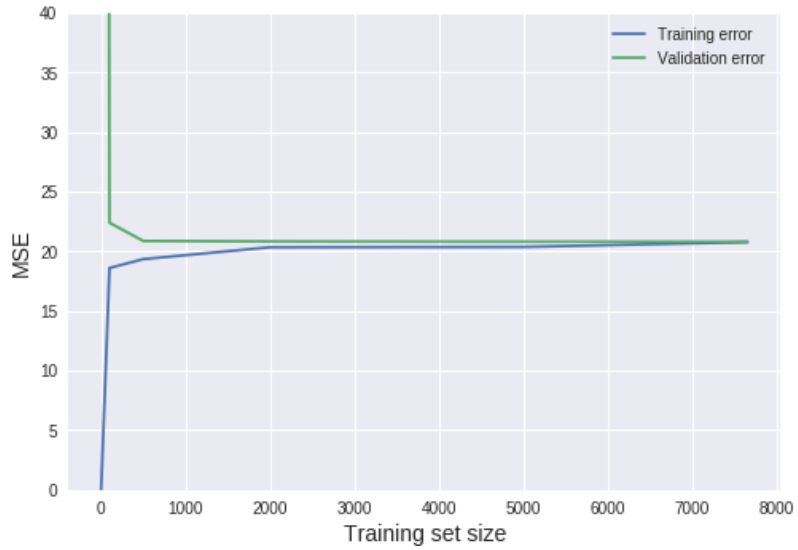
Slično unakrsnoj validaciji, LOOCV iterativno koristi samo jedan podatak za validiranje algoritma te ostatak podataka za učenje algoritma.

Napomenimo da je konvencija koristiti Train & Test metodu te unakrsnu validaciju zajedno. Konkretnije, najprije podijelimo podatke na train i test skupove, a zatim koristimo unakrsnu validaciju na skupu za treniranje putem kojeg bolje možemo naučiti potrebne parametre modela, a onda validiramo cjelokupni algoritam na testnom skupu.

Nadalje, osim spomenutih tehnika, postoje i razne mjere kojima općenito dobivamo uvid u model. Takve mjere se baziraju na statističkoj procjeni rezultata algoritma. U nastavku ćemo pojasniti neke od tih metoda.

Krivulja točnosti učenja

Krivulja učenja je krivulja koja nam prikazuje napredak (smanjenje greške na testnom skupu) algoritma kroz njegove iteracije. Alternativa takvoj krivulji je da se umjesto iteracija algoritma, gleda broj podataka na kojemu je algoritam treniran.



Matrica konfuzije

Matrica konfuzije je tablica koja se koristi radi evaluacije učinkovitosti klasifikacijskog mo-

dela.

		Stvarna klasa	
		Pozitivni	Negativni
Predviđeno modelom	Pozitivni	TP	FP
	Negativni	FN	TN

TP- true positives (broj stvarno pozitivnih primjera, točno predviđenih od strane modela)
FP- false positives (broj stvarno negativnih primjera, koji su netočno predviđeni od strane modela kao pozitivni)
TN- true negatives (broj stvarno negativnih primjera, koji su točno predviđeni od strane modela kao negativni)
FN- false negatives (broj stvarno pozitivnih primjera, koji su netočno predviđeni od strane modela kao negativni)

Iz matrice konfuzije proizlaze mnoge mjere koje nam daju bolju intuiciju o prediktivnoj moći algoritma. Omjer točno klasificiranih primjera u odnosu na ukupan broj primjera:

$$\text{Točnost (en. Accuracy)} = \frac{TP+TN}{TP+FP+TN+FN} \times 100$$

nam nam daje postotak točnih predviđanja našeg algoritma.
Omjer pozitivnih primjera koje je model prepoznao kao pozitivne, od ukupnog broja pozitivnih primjera:

$$\text{Osjetljivost (en. Sensitivity/Recall/True positive rate)} = \frac{TP}{TP+FN}$$

nam intuitivno govori koliko je algoritam dobar u pronalasku pozitivnih primjera. Inherentno, govori nam koliko veliku grešku druge vrste činimo, odnosno grešku da pozitivne primjere predvidimo negativnima. Algoritam koji ima slabu osjetljivost bi često klasificirao primejre kao negativnima s ciljem da ne bude ”pretjerano pozitivan”.
Omjer dobro prepoznatih negativnih primjera, od ukupnog broja negativnih primjera:

$$\text{Specifičnost (en. Specificity)} = \frac{TN}{FP+TN}$$

nam intuitivno govori koliko dobro raspoznajemo negativne primjere. Inherentno, govori nam koliko veliku grešku prve vrste činimo, odnosno grešku da negativne primjere predivimo kao

pozitivne. Algoritam koji ima slabu specifičnost bi često klasificirao podatke kao pozitivne u nadi da pronađe pozitivan primjer. Intuitivno, osjetljivost i specifičnost nam opisuju koliko dobro algoritam razlikuje odnosno diskriminira između pozitivne i negativne klase.
Omjer točno pozitivnih primjera koje je model prepoznao kao pozitivne, od ukupnog broja pozitivno predviđenih primjera:

$$\text{Preciznost (en. Precision)} = \frac{TP}{TP+FP}$$

nam govori koliko algoritam dobro predviđa nad pozitivnom klasom. Općenito, preciznost pada kako osjetljivost raste jer što više pozitivne primjere previđamo pozitivnima, to smo manje precizni nad pozitivnom klasom. Stoga bi htjeli imati dobar balans među njima. Zbog toga uvodimo mjeru:

$$F_\beta = \frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P+R}$$

gdje je P Preciznost a R Osjetljivost algoritma a β relativna važnost preciznosti u odnosu na osjetljivost. Drugi razlog ovakve mjere je da sažmemo mjere preciznosti i osjetljivosti u jedno, te si olakšamo usporedbu raznih modela.
Uobičajeno se koristi mjera

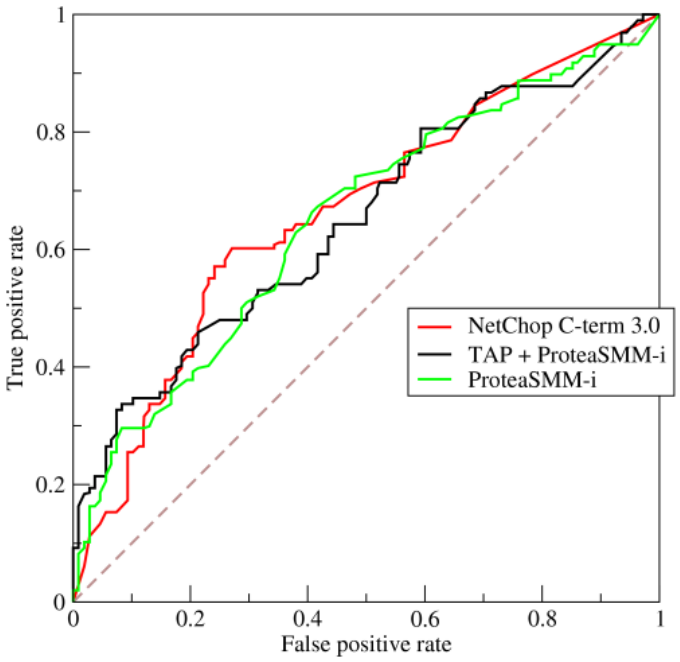
$$F_1 = \frac{2PR}{P+R}$$

koja predstavlja harmoničku sredinu P i R i ponaša se kao aritmetička sredina kada su P i R relativno blizu. Kada P i R nisu blizu, tada je harmonička sredina manja od aritmetičke što odražava činjenicu da želimo dobar balans između preciznosti i osjetljivosti. Želimo da takva mjera bude što bliža 1 jer za tu vrijednost imamo najveću preciznost i osjetljivost algoritma.

ROC krivulja

Krivulja koja prikazuje odnos TPR (True positive rate) odnosno broj korektnih klasifikacija u (pozitivnoj) klasi u odnosu na ukupan broj pozitivnih primjera i FPR (False positive rate) odnosno broj krivih klasifikacija u (pozitivnoj) klasi u odnosu na ukupan broj negativnih primjera zove se ROC (en. Reciever Operating Characteristic) krivulja.

$$TPR = \text{Osjetljivost} = \frac{TP}{TP+FN}$$
$$FPR = 1 - \text{Specifičnost} = \frac{FP}{FP+TN}$$



Kako mijenjamo prag klasifikacije između pozitivnih i negativnih primjera za algoritam,tako dobivamo različite omjere TPR-a i FPR-a te upravo njih možemo prikazati na grafu. Najbolja moguća predikcija bi se nalazila u gornje ljevom kutu gdje bi imali savršen klasifikator sa osjetljivosti i specifičnosti jednakima jedan. Simetrala kvadranta predstavlja beznačajan algoritam koji se mogao simulirati slučajnim algoritmom.

Površina ispod ROC krivulje

Površina ispod ROC krivulje (en. Area Under ROC curve, AUROC) nam daje vjerojatnost da algoritam nasumično izabran pozitivan primjer rangira iznad (odnosno ima veću vjerojatnost da bude klasificiran kao pozitivan) nasumično odabranog negativnog primjera.
Takva površina korisna mjera jer je invarijantna obzirom na prag klasifikacije algoritma.

4 Eksperiment

4.1 Podatkovni skup

Originalni podatkovni skup sastoji se od financijskih izvještaja za 80-ak tisuća poduzeća u Hrvatskoj od 2007 do 2014 godine. Financijska izvješća se sastoje od bilance, računa dobiti i gubitka te novčanog toka. Ujedno su za svako poduzeće specificirani razni opisni parametri kao što su sektor u koji se poduzeće svrstava, županija poduzeća, veličina poduzća, te mnogi

drugi.
Kao što smo napomenuli ranije, bavimo se s 5 različitih problema za predikciju ulaska u predstečaj u ovisnosti o 5 duljina perioda. Analizirajmo zato podatkovni skup kada je cilj predvidjeti predstečaj poduzeća u razmaku od 5 godina.

Takav skup se sastoji od 264154 različitih poduzeća od kojih je samo 641 završilo u predstečaju, dok 263513 nije. Možemo primjetiti veliku asimetričnost podataka. To je konkretan problem na koji ćemo morati obratiti pozornost kod izvođenje te evaluiranje modela. Iz analize podataka možemo primjetiti da je najviše podataka u sektoru "Usluge popravaka" nakon čega slijede "usluge izajmljivanja". U analizi ćemo se fokusirati na par najvećih sektora, dok se kasnije rad može potencijalno proširiti na sve dostupne sektore.

4.2 Eksperimentalni setup

Cilj našeg eksperimenta je evaluirati model te ga usporediti s baseline modelima logističke regresije, stabla odluke i modelom slučajne šume na način opisan u sljedećem dijelu na podacima o hrvatskim poduzećima.

4.3 Evaluacija modela

Kao što je već objašnjeno u poglavlju Podatkovni skup, podaci su jako asimetrični. Postoji mnogo više poduzeća koja nisu završila u predstečajnoj nagodbi. Stoga, ako želimo evaluirati razne modele nad ovakim problemom, potrebna je metrika koja je indiferentna na asimetričnosti podataka.
ROC krivulja, te inherentno površina ispod ROC krivulje, je takva metrika, dok F_1 nije. Raspišimo sada TPR i FPR.

$$TPR = \text{Osjetljivost} = \frac{TP}{TP+FN}$$
$$FPR = 1 - \text{Specifičnost} = \frac{FP}{FP+TN}$$

Iz TPR vidimo da ako povećamo pozitivne primjere za neki faktor, povećali bi se i TP i FN, što znači da ne bi došlo do promjene TPR-a. Analogno, iz FPR-a vidimo da ako povećamo negativne primjere za neki faktor, FPR se nebi mjenjao. Time vidimo da je površina indiferentna asimetričnim podacima.

Osim ovakve evaluacije, koristit ćemo train i test metodu zajedno sa unakrsnom validacijom. Opišimo malo kako takve tehnike probira i AUROC surađuju.

Prvi korak je podijeliti podatke na skup za treniranje i skup za testiranje. Zatim, koristimo unakrsnu validaciju na skupu za treniranje radi poboljšanja parametara algoritma. Konkretnije koristimo deseterostruku unakrsnu validaciju sa AUROC evaluacijom. To znači da podijelimo skup za treniranje na 10 jednakih dijelova te iterativno treniramo na 9 skupova dok na preostalom skupu validiramo algoritam. Upravo za tu validaciju koristimo AUROC mjeru. To znači da na kraju postupka unakrsne validacije imamo 10 AUROC mjera koje ćemo uprosječiti. Verzija algoritma sa najvećom prosječnom AUROC mjerom biti će izabrani algoritam koji ćemo onda evaluirati na testnom skupu koji smo dobili na početku. Pritom ćemo opet koristiti AUROC mjeru. Opisani postupak izvodit će se za svaki par hiperparametara algoritma koji prolazi nekom diskretnom predodređenom mrežom vrijednosti. Na kraju ćemo izabrati finalni algoritam koji će imati najveću AUROC mjeru.

4.4 Interpretacija

Jedan od razloga odabira EXGB modela je njegova interpretabilnost. Nakon prilagodbe modela razmatranjem njegove strukture možemo uvidjeti koje ekonomske mjere bitno sudjeluju pri donošenju predviđanja unutar modela. U [1] gledali su učestalost pojavljivanja određene značajke u stablima ansambla kao indikator te važnosti i tako došli do zaključka na koje ekonomske indikatore bi bilo dobro obratiti pozornost pri analizi poduzeća u svrhu predviđanja propadanja. Planiramo na taj način analizirati dobiveni model te usporediti zaključke s "tradicionalnim" ekonomskim znanjem. Zanimljivo bi bilo čuti mišljenje ekonomskih stručnjaka i ljudi koji se u praksi bave ovim problemom o dobivenim rezultatima.

5 Literatura

[1] Maciej Zieba, Sebastian K. Tomczak, Jakub M. Tomczak: *Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction*, Expert Systems with Applications, str: 93-101

[2] Esteban Alfaro, Noelia Garcia, Matias Gamez, David Elizondo: *Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks*, Decision Support Systems, str: 110-122

[3] Mohamed A. Rujoub, Doris M. Cook, Leon E. Hay: *Using Cash Flow Ratios To Predict Business Failures*, Journal of Managerial Issues, str: 75-90

[4] Deron Liang, Chia-Chi Lub, Chih-Fong Tsaic, Guan-An Shiha: *Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study*, European Journal of Operational Research, str: 561-572

[5] Loris Nanni, Alessandra Lumini: *An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring*, Expert Systems with Applications journal, str: 3028-3033