Predviđanje ulaska poduzeća u Hrvatskoj u postupak predstečajne nagodbe

David Bojanić
Matematički odsjek
Prirodoslovno-matematički fakultet
Sveučilište u Zagrebu
david.bojanic4@gmail.com

Domagoj Demeterfi
Matematički odsjek
Prirodoslovno-matematički fakultet
Sveučilište u Zagrebu
dodemet@student.math.hr

***Sažetak*—Predviđanje propadanja poduzeća je područje od velikog interesa u ekonomiji. Cilj je razviti prediktivni model koji koristeći ekonometrijske mjere poduzeća predviđa financijsko stanje poduzeća u budućnosti. Za razvoj takvog modela u ovom radu koristimo relativno novu tehniku ekstremnog gradijentnog boostinga za učenje ansambla stabala. Za ocjenu uspješnosti našeg pristupa koristimo prikupljene podatke o poduzećima u Hrvatskoj u periodu između 2007. i 2014. godine te ga uspoređujemo s nekim referentnim metodama korištenim u ovom području. Naposljetku dobiveni model interpetiramo s ekonomskog stajališta.**

I. Uvod

Cilj predviđanja propadanja poduzeća je procijeniti financijsko stanje i perspektivu poduzeća u budućnosti. To je iznimno važno 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 nije čudno što taj problem već dugo privlači pozornost istraživača.

A. 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 [10] primijenili ansambl boosted stabala sa sintetičkim značajkama u svrhu predviđanja stečaja poljskih poduzeća. Svoj model su eksperimentom usporedili s pristupima uobičajenim u području predviđanja stečaja te dobili značajno bolje rezultate. Za opis poduzeća oni koriste 64 financijska omjera često korištenih u integriranim financijskim modelima i financijskoj analizi. Osim spomenutih omjera koriste i tzv. sintetičke 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 [1] 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 temeljenih na računovodstvenim podacima, ali i neke kvalitativne varijable kao što su veličina firme, sektor i vlasnička struktura. U okviru neuronske mreže analizirali su relativni doprinos svake input varijable globalnim svojstvima pomoću analize osjetljivosti. 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 (također redom po važnosti): ekonomska profitabilnost (pokazuje uspješnost firme u korištenju imovine), razina zaduženosti, efikasnost i varijabla koja procjenjuje veličinu firme.

Rujoub et al. u [8] ocjenjuju korisnost varijabli vezanih uz novčani tok u predviđanju stečaja. Multivarijatnom diskriminantnom analizom ispituju hipoteze o diskriminantnoj sposobnost podataka o novčanom toku (u obliku financijskih omjera). Testiranjem tih hipoteza došli su do zaključka 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 povećava moć predviđanja kasnijih.

Liang et al. u [6] ispituju diskriminatornu moć dobivenu kombiniranjem različitih kategorija financijskih omjera (FR) i indikatora korporativnog upravljanja (CGI). Preciznije, uzeli su u obzir sedam FR kategorija (koeficijenti profitabilnosti, kapitalne strukture, obrtaja, novčanog toka te rasta) i pet CGI kategorija (struktura upravnog odbora, vlasnička struktura, prava od novčanih tokova te zadržavanje ključnog osoblja). Da bi odredili najbolju kombinaciju FR-ova i CGI-ova 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, ali napominju kako korisnost CGI značajki ovisi o tržištu (kao razloge tome navode drugačije definicije stečaja na tržištima te da je opseg u kojemu su CGI-ovi povezani s karakteristikama kompanije ovisan o tržištu). Nadalje, zaključuju da su najvažnije značajke za predviđanje stečaja FR kategorije solventnost i profitabilnost te CGI kategorije struktura upravnog odbora i vlasnička struktura.

B. Sadržaj rada

Obično se u literaturi koja se bavi predviđanjem propadanja poduzeća predviđa ulazak poduzeća u stečaj. Mi smo kroz razgovor sa stručnjakinjom 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. Zanimljivo bi u budućem istraživanju bilo predviđati i trajanje predstečajne nagodbe i/ili iznos tražbine nakon prijeboja u predstečajnoj nagodbi.

U radu postavljamo model za predviđanje ulaska poduzeća u postupak predstečajne nagodbe. Pokazano je da se anambli klasifikatora mogu uspješno primijeniti za predviđanje stečaja ([7]) te su značajno bolji nego druge metode ([1]). Nedavno je boosting metoda modificirana tako da se optimizira Taylorov razvoj funkcije gubitka, pristup poznat kao ekstremni gradijentni boosting (en. Extreme Gradient Boosting). Taj pristup vrlo je uspješno primijenjen za rješavanje mnogih klasifikacijskih problema pa tako i za predviđanje stečaja ([10]). Nadalje, model nije osjetljiv na problem neuravnoteženosti podataka (jedan od značajnih problema ovog područja) jer dopušta korištenje površine ispod krivulje (en. Area Under urve, AUC) za evaluaciju te primorava pravilan poredak neuravnoteženih podataka. Stoga smo za osnovu našeg rada odabrali model ansambla regulariziranih boosted stabala koji je treniran tehnikom ekstremnog gradijentnog boostinga. Također smo testirali prethodni model uz dodatno korištenje sintetičkih značajki (model predložen u [10]). Sintetičke značajke računaju se u svakom koraku boostinga kombinirajući postojeće značajke nekom od osnovnih aritmetičkih operacija $(+, -, *, /)$. 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.

Rješenje testiramo na podacima o hrvatskim poduzećima u građevinskom sektoru u periodu od 2007. do 2014. godine. Konkretnije, gore navedenim modelom rješavamo 5 problema binarne klasifikaciju ovisnosti o periodu predvidanja: poduzeće je ušlo u postupak predstečajne nagodbe nakon jedne, dvije, tri, četiri ili pet godina.

U području predviđanja propadanja poduzeća osim same točnosti modela druga važna komponenta je interpretabilnost. Razumljivi modeli nam dopuštaju uvid u proces donošenja odluke unutar modela što može pomoći u razumijevanju razloga propadanja poduzeća. To nadalje može pomoći u poduzimanju koraka s ciljem sprječavanja istog. Stoga osim samog razvoja i testiranja modela predviđanja ulaska u postupak predstečajne nagodbe u radu se bavimo i interpretacijom dobivenog modela s ekonomskog stajališta.

C. Organizacija rada

U dijelu II uvodimo ansambl boosted stabala kao model za predviđanje ulaska poduzeća u postupak predstečajne nagodbe te opisujemo korištenje sintetičkih značajki. U dijelu III opisujemo provedeni eksperiment i prezentiramo njegove rezultate ostvarene na podatkovnom skupu o hrvatskim poduzećima te intepretiramo dobiveni model. Mogućnosti daljnjeg istraživanja dane su u dijelu IV.

II. Metodologija

A. Extreme gradient boosting

Označimo s $x \in \mathcal{X}$ vektor značajki koje opisuju poduzeće, gdje je $\mathcal{X} \subseteq \mathbb{R}^p$ prostor značajki i s $y \in \{0, 1\}$ oznaku koja nam govori je li poduzeće završilo u postupku predstečajne nagodbe $(y = 1)$ ili ne $(y = 0)$.

Modeli stabla odluke dijele prostor značajki $\mathcal{X}$ u skup od T pravokutnih, nepreklapajućih dijelova $R_1, ..., R_T$ te na svakom od tih dijelova prilagođavaju jednostavan model poput konstante. Stoga model stabla možemo zapisati u obliku

$$f(x) = \sum_{j=1}^{T} w_j I(x \in R_j), \tag{1}$$

gdje je w_j težina na listu koji odgovara dijelu R_j.

Želimo naučiti ansambl M stabala koji možemo zapisati kao

$$f^{(M)}(x) = \sum_{m=1}^{M} f_m(x), \tag{2}$$

gdje su f_m stabla odluke oblika (1). Da bi donijeli odluku za novi primjer x možemo izračunati uvjetnu vjerojatnost

$$p(y = 1|x) = \sigma(f(x)), \tag{3}$$

gdje je $\sigma(a) = \frac{1}{1-\exp(-a)}$ sigmoid funkcija.

Za trening skup $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ model treniramo minimizirajući kriterij

$$J_M(f^{(M)}) = L(f^{(M)}) + \Omega(f^{(M)})$$
$$= \sum_{i=1}^{n} l(y_i, f^{(M)}(x_i)) + \sum_{m=1}^{M} \Omega(f_m), \tag{4}$$

gdje je $L(f^{(M)}) = \sum_{i=1}^n l(y_i, f^{(M)}(x_i))$ funkcija gubitka, a $\Omega(f^{(M)}) = \sum_{m=1}^M \Omega(f_m)$ (aditivni) član regularizacije. U ovom radu za problem binarne klasifikacije $(y_i \in \{0, 1\})$ koristimo logistički gubitak, tj.

$$l(y_i, f^{(m)}(x_i)) = y_i \log(1 + \exp(-f^{(m)}(x_i)))$$
$$+(1 - y_i) \log(1 + \exp(f^{(m)}(x_i))). \tag{5}$$

Boosting metode za stabla sekvencijalno u svakoj iteraciji dodaju trenutnom modelu stablo koje poboljšava fit cjelokupnog modela. Pretpostavimo da smo naučili $f^{(m-1)} = \{f_1, ..., f_{m-1}\}$. U m-toj iteraciji treniranja ansamblu dodajemo novo stablo $f_m(x) = \sum_{j=1}^T w_{jm} I(x \in R_{jm})$. Prvo primijetimo da funkciju gubitka za jedan primjer možemo zapisati kao

$$l(y_i, f^{(m)}(\mathbf{x}_i)) = l(y_i, f^{(m-1)}(\mathbf{x}_i) + f_m(\mathbf{x}_i)), \tag{6}$$

a član regularizacije (zbog pretpostavke aditivnosti) kao

$$\sum_{j=1}^{m} \Omega(f^{(j)}) = \Omega(f_m) + \Omega(f^{(m-1)}) = \Omega(f_m) + const. \tag{7}$$

($\Omega(f^{(m-1)}) = const.$ jer miniziramo kriterij s obzirom na f_m). Uzimajući u obzir (6) i (7) kriterij učenja (4) možemo zapisati kao

$$J_m(f^{(m)}) = \sum_{i=1}^{n} l(y_i, f^{(m-1)}(\mathbf{x}_i) + f_m(\mathbf{x}_i)) + \Omega(f_m) + const. \tag{8}$$

U metodi ekstremnog gradijentnog boostinga gornji kriterij (8) aproksimira se s (Taylorov razvoj l oko $f^{(m-1)}(x_i)$)

$$J_m(f^{(m)}) = \sum_{i=1}^{n} [l(y_i, f^{(m-1)}(x_i)) + g_m(x_i) f_m(x_i)$$
$$+\frac{1}{2} h_m(x_i) f_m(x_i)^2] + \Omega(f_m) + const. \tag{9}$$

gdje je

$$g_m(x_i) = \frac{\delta L(y_i, f^{(m-1)}(x_i))}{\delta f^{(m-1)}(x_i)} \tag{10}$$

te

$$h_m(x_i) = \frac{\delta^2 L(y_i, f^{(m-1)}(x_i))}{\delta f^{(m-1)}(x_i)^2}. \tag{11}$$

Uzimajući u obzir da je l logistički gubitak (5) može se pokazati

$$g_m(x_i) = \sigma(f^{(m-1)}(x_i)) - y_i, \tag{12}$$

$$h_m(x_i) = \sigma(f^{(m-1)}(x_i))(1 - \sigma(f^{(m-1)}(x_i))). \tag{13}$$

Zapišimo aproksimaciju kriterija (9) uzimajući u obzir (1) (f_m je stablo)

$$J_m(f^{(m)}) = \sum_{i=1}^{n} [g_m(x_i) \sum_{j=1}^{T} w_{jm} I(x_i \in R_{jm})$$
$$+\frac{1}{2} h_m(x_i) (\sum_{j=1}^{T} w_{jm} I(x_i \in R_{jm}))^2] + \Omega(f_m) + const. \tag{14}$$

Pravokutnici $R_{jm}, j = 1, ..., T$ su disjunktni pa dalje imamo

$$J_m(f^{(m)}) = \sum_{i=1}^{n} [g_m(x_i) \sum_{j=1}^{T} w_{jm} I(x_i \in R_{jm})$$
$$+\frac{1}{2} h_m(x_i) \sum_{j=1}^{T} w_{jm}^2 I(x_i \in R_{jm})] + \Omega(f_m) + const.$$
$$= \sum_{j=1}^{T} \sum_{i \in I_{jm}} \left[g_m(x_i) w_{jm} + \frac{1}{2} h_m(x_i) w_{jm}^2 \right] + \Omega(f_m) + const. \tag{15}$$

gdje je I_{jm} skup indeksa i za koje $x_i \in R_{jm}$. Stavimo $G_{jm} = \sum_{i \in I_{jm}} g_m(x_i)$ i $H_{jm} = \sum_{i \in I_{jm}} h_m(x_i)$ pa imamo

$$J_m(f^{(m)}) = \sum_{j=1}^{T} \left[G_{jm} w_{jm} + \frac{1}{2} H_{jm} w_{jm}^2 \right] . + \Omega(f_m) + const. \tag{16}$$

Postoje različite mogućnosti regularizacije no ovdje pretpostavljamo da je član regularizacije oblika

$$\Omega(f_m) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_{jm}^2, \tag{17}$$

gdje su γ i λ hiperparametri (parametri regularizacije). Dakle, penaliziramo broj čvorova stabla te težine listova. Uz regularizaciju oblika (17) aproksimacija kriterija (16) prelazi u

$$J_m(f^{(m)}) = \sum_{j=1}^{T} \left[G_{jm} w_{jm} + \frac{1}{2} H_{jm} w_{jm}^2 \right]$$
$$+\gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_{jm}^2 + const.$$
$$= \sum_{j=1}^{T} \left[G_{jm} w_{jm} + \frac{1}{2} (H_{jm} + \lambda) w_{jm}^2 \right] . + \gamma T + const. \tag{18}$$

Uz poznatu (fiksnu) strukturu stabla (tj. pravokutnike $R_{jm}, j = 1, ..., T$) optimalne težine su dane s

$$\tilde{w}_{jm} = -\frac{G_{jm}}{H_{jm} + \lambda}, \tag{19}$$

a vrijednost kriterija za te težine tada je

$$J_m(\tilde{f}^{(m)}) = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_{jm}^2}{H_{jm}+\lambda} + \gamma T + const. \tag{20}$$

Učenje strukture stabla zapravo je traženje podjela. Krenemo od korijena stabla te tražimo najbolju značajku i najbolju vrijednost za podjelu u korijenu. Traženje podjela dalje nastavljamo do nekog kriterija zaustavljanja. Kriterij po kojem tražimo podjele je porast informacije (en. information gain)

$$InfoGain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma, \tag{21}$$

gdje je $\frac{G_L^2}{H_L+\lambda}$ gubitak u lijevom djetetu, $\frac{G_R^2}{H_R+\lambda}$ gubitak u desnom djetetu, a $\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$ gubitak u čvoru u kojem razmatramo podjelu ako se podjela ne ostvari. $InfoGain$ je zapravo smanjenje funkcije gubitka odabranom podjelom.

Ukratko, algoritam učenja ansambla se može opisati na sljedeći način. U svakoj iteraciji algoritma m odredimo strukturu stabla $\hat{f}_m$ kojeg dodajemo postojećem modelu $\hat{f}^{(m-1)}$, tj. odredimo $\hat{R}_{jm}, j = 1, ..., T$. Zatim odredimo težine $\hat{w}_{jm}, j = 1, ..., T$ za tu strukturu. Stavimo $\hat{f}_m(x) = \eta\sum_{j=1}^{T}\hat{w}_{jm}I(x \in \hat{R}_{jm})$ te $\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \hat{f}_m(x)$, gdje je $0 < \eta \leq 1$ hiperparametar, tzv. parametar sažimanja ili stopa učenja.

B. Sintetičke značajke

Zieba et al. u [10] predlažu korištenje dodatnih sintetičkih značajki u svrhu poboljšanja predikcijskih svojstava ansambla stabala.

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. Konkretnije, definiramo kategorijsku distribuciju $\boldsymbol{\theta} = \left(\theta^{(1)}, ..., \theta^{(d)}, ..., \theta^{(D)}\right)$, tj. vektor vjerojatnosti odabira pojedine značajke, gdje

$$\theta^{(d)} = \frac{m_d}{\sum_{i=1}^{D} m_i},$$

a m_d je broj pojavljivanja d-te značajke u stablima ansambla. Posljedica ovog pristupa je da će najpopularnije značajke biti odabrane za stvaranje novih. Na to možemo gledati kao na evolucijski pristup koji odabire "najsnažnije" roditelje značajke za potomak značajku. Operacija se uniformno bira iz skupa $\{+, -, *, /\}$.

Postupak stvaranja ansambla uz generiranje sintetičkih značajki u svakom koraku boostinga opisan je s Algoritmom 1. U svakoj iteraciji učenja t postojećem modelu $\left\{f^{(1)}, ..., f^{(t-1)}\right\}$ dodajemo stablo $f^{(t)}$ koje učimo tehnikom ekstremnog gradijentnog boostinga koristeći trening skup $\mathcal{D}$.

U ovisnosti o važnostima značajki m_d, $d = 1, ..., D$ izračunatim iz dosad poznatog modela $\left\{f^{(1)}, ..., f^{(t)}\right\}$ nastavljamo učenje samo s onim značajkama za koje je m_d veće od praga ϵ. Dosad naučen model dalje koristimo da bi odredili popularnost značajki, tj. da procijenimo distribuciju $\boldsymbol{\theta}$.

Sintetičke značajke generiraju se na sljedeći način. Biramo dvije značajke f_1 i f_2 iz distribucije $\boldsymbol{\theta}$ te uniformno biramo operaciju $\circ$ iz skupa $\{+, -, *, /\}$. Nakon toga vrijednost nove značajke $f_{new} = f_1 \circ f_2$ računa se za sve primjere u trening skupu $\mathcal{D}$ i on se proširuje njome. Postupak stvaranja sintetičkih značajki nastavlja se sve dok ne postignemo željeni broj sintetičkih značajki D_{new}. Na proširenom trening skupu dalje učimo stablo $f^{(t+1)}$.

Input: trening skup $\mathcal{D}$, broj sintetičkih značajki D_{new}, broj stabala u ansamblu T, prag prihvaćanja značajki ϵ
Output: skup stabala koja su u ansamblu $\left\{f^{(1)}, ..., f^{(T)}\right\}$

```
for t = 1, ..., T do
    Nauči f^(t) koristeći D tehnikom ekstremnog gradijentnog boostinga;
    Ukloni iz D značajke za koje je m_d < ε;
    Procijeni θ iz ansambla {f^(1), ..., f^(t)};
    for d = 1, ..., D_new do
        Izaberi značajke f_1 i f_2 iz distribucije θ;
        Izaberi operaciju ∘ uniformno iz skupa {+, −, *, /};
        Generiraj novu značajku f_new = f_1 ∘ f_2;
        Proširi D s f_new;
    end
end
return {f^(1), ..., f^(T)};
```

Algoritam 1: Učenje ansambla boosted stabala tehnikom ekstremnog dradijentnog boostinga uz korištenje sintetičkih značajki

III. Eksperiment

A. Podatkovni skup

S ciljem evaluacije kvalitete našeg pristupa prikupili smo podatke o hrvatskim poduzećima. Podatke smo prikupili iz financijskih izvještaja te izvještaja o dodatnim statističkim podacima za period od 2007. do 2014. godine. Financijska izvješća se sastoje od bilance, računa dobiti i gubitka te novčanog toka. Podaci s informacijom kada je poduzeće završilo predstečajnoj nagodbi dostupni su u originalnom podatkovnom skupu za poduzeća koja su završila u predstečaju u preiodu od 2012. do 2014. godine. Podatkovni skup jako je neuravnotežen,

postoji puno više poduzeća koja nisu završila u postupku predstečajne nagodbe nego onih koji jesu.

Proces odabira podataka koje ćemo koristiti za učenje i evaluaciju sastoji se od specificiranja perioda i sektora u kojem promatramo poduzeća te odabira značajki.

Odabrali smo sektor građevine jer se u promatranom periodu (oko krize 2008.) mnogo građevinskih firmi našlo u problemima.

Nismo radili a priori selekciju varijabli jer je ona ugrađena u model s kojim pristupamo problemu. Iz opisa ekstremnog gradijentnog boostinga izloženog u II vidimo da istovremeno učimo model i radimo selekciju varijablu. Dakle, za značajke koje opisuju poduzeće unutar modela odabrali smo sve dostupne: sve stavke iz bilance, računa dobiti i gubitka, novčanog toka poduzeća te sve stavke izvještaja o dodatnim statističkim podacima. Broj navedenih značajki je 380, a za točan popis svih značajki možete se javiti autorima.

Za sektor građevine razlikujemo 5 problema klasifikacije u ovisnosti o 5 različitih perioda predviđanja. Na primjer, za klasifikacijski problem predviđanja ulaska u postupak predstečajne nagodbe nakon 5 godina uzeli smo značajke o poduzeću iz financijskih izvješća iz određene godine i odgovarajuću oznaku koja nam govori je li poduzeće završilo u predstečaju nakon 5 godina. Konkretnije, s trenutno dostupnim podacima za ovaj problem gledali smo financijske izvještaje iz 2007., 2008. i 2009. godine te je li poduzeće završilo u predstečaju u 2012., 2013. i 2014. godini, respektivno. Broj poduzeća koje jesu i nisu završila u postupku predstečajne nagodbe po podatkovnim skupovima za ovih 5 problema je

- *Nakon1Godine* – 301 u predstečaju, 29358 nije (ukupno 29659 poduzeća)
- *Nakon2Godine* – 153 u predstečaju, 29625 nije (ukupno 29778 poduzeća)
- *Nakon3Godine* – 151 u predstečaju, 30231 nije (ukupno 30382) poduzeća
- *Nakon4Godine* – 153 u predstečaju, 29808 nije (ukupno 29961 poduzeća)
- *Nakon5Godina* – 129 u predstečaju, 28549 nije (ukupno 28678 poduzeća).

Vidimo da su uzorci za svih 5 problema jako neuravnoteženi, no primijetimo da je za prvi problem omjer broja poduzeća koja jesu i nisu u predstečaju oko dva puta veći nego isti omjer kod ostalih problema.

B. Postavke eksperimenta

Cilj eksperimenta je evaluirati pristupe XGB (ansambl stabala treniran tehnikom ekstremnog gradijentnog boostinga) i XGBSZ (XGB model uz dodatno korištenje sintetičkih značajki) na podacima o hrvatskim poduzećima opisanim u prethodnom dijelu. Također, usporediti ćemo ih s klasifikacijskim metodama logističke regresije (LR), stabla odluke (DT) i metodom slučajne šume (RF) koje su često korištene u području predviđanja propadanja poduzeća. Način evaluacije i usporedbe opisan je u nastavku.

Kao što smo vidjeli u prethdnom dijelu podaci su neuravnoteženi, tj. postoji mnogo više poduzeća koja nisu završila u predstečajnoj nagodbi. Stoga ako želimo evaluirati modele nad ovakvim podacima, potrebna je mjera koja nije osjetljiva na neuravnoteženost podataka.

ROC (en. receiver operating characteristic) krivulja ima takvo svojstvo - nije osjetljiva na promjenu u distribuciji klasa. Ako se proporcija pozitivnih i negativnih primjera u testnom skupu promijeni ROC krivulja se neće promijeniti. To vidimo iz matricu konfuzije. Distribucija klasa (omjer pozitivnih i negativnih primjera) je zapravo odnos lijevog i desnog stupca matrice konfuzije. Stoga će svaka mjera uspješnosti koja koristi vrijednosti iz oba stupca biti osjetljiva na neuravnoteženost klasa. Takve mjere su npr. točnost, preciznost i F-mjera pa će promjena u distribuciji klasa promijeniti i vrijednosti tih mjera iako se fundamentalna sposobnost predviđanja ne mijenja. ROC krivulje se temelje na TPR (en. true positive rate) i FPR (en. false positive rate) koji proizlaze iz jednog stupca pa stoga ne ovise o distribuciji klasa. Iz ovoga možemo zaključiti da ni površina ispod ROC krivulje (en. Area Under ROC Curve, AUROC) nije osjetljiva na neuravnotežene podatke.

Još jedan razlog zbog kojeg je AUROC mjera primjerena za područje predviđanja propadanja poduzeća je to što različite vrste greške (FP greške i FN greške) ovdje imaju različite troškove ili posljedice.

Konkretno za svaki od 5 podatkovnih skupova (za 5 problema klasifikacije), u prvom koraku koristimo Train & Test metodu, tj. podijelili smo podatke na skup za treniranje i skup za testiranje. Zatim koristimo deseterostruku unakrsnu validaciju s AUROC mjerom na skupu za treniranje. Nakon toga provest ćemo i dodatnu ocjenu (također AUROC mjerom) na spomenutom izoliranom testnom skupu.

C. Rezultati

Rezultati eksperimenta prikazani su u tablici I (prosječni AUROC za desetorostruku unakrsnu validaciju kao što je opisano u prethodnom dijelu). Parametri su osnovnim modela ostavljeni defaultno zadani. XGB model treniran je sa 100 koraka boostinga. Ostali parametri XGB modela dobiveni su metodom pretraživanja rešetke. XGBSZ model treniran je sa 30 koraka boostinga uz generiranje 30 sintetičih značajki u svakom koraku. Parametri XGBSZ modela dobiveni su "ručnim" testiranjem različitih parametara uz mjeru uspješnosti AUROC uz desetorostruku unakrsnu validaciju.

Iz tablice I može se uočiti da je XGB model bolji od ostalih metoda (pod ovim uvjetima, uz mjeru dobrote AUROC) na podatkovnim skupovima za svih 5 problema. XGBSZ radi najlošije što nije baš očekivano. Jedan od mogućih razloga u tome je što se u financijama kao prediktori propadanja poduzeća često koriste omjeri

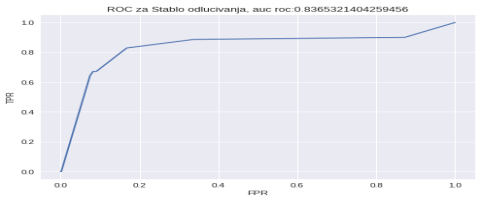
	1 god.	2 god.	3 god.	4 god.	5 god.
LR	0.700	0.628	0.562	0.630	0.674
DT	0.852	0.792	0.766	0.719	0.705
RF	0.899	0.844	0.806	0.785	0.790
XGB	0.935	0.878	0.817	0.811	0.844
XGBSZ	0.636	0.573	0.562	0.573	0.550

Tablica I

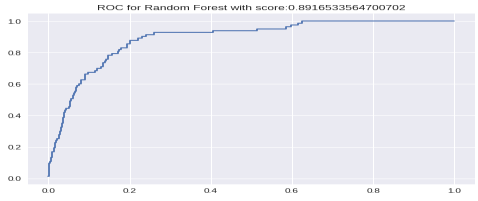
stavki iz financijskog izvještaja. Zbog manjka preprocesiranja u podatkovnom skupu nalazi se puno nula što je onemogućilo da se stvaraju sintetičke značajke s operacijom djeljenja.

Nakon ovoga istrenirali smo modele na cijelom trening skupu za problem previđanja jedne godine unaprijed i ocijenili grešku na prije izdvojenom testnom skupu za isti problem. ROC krivulje s odgovarajućim AUROC mjerama za dobivene modele mogu se vidjet na slikama ispod.

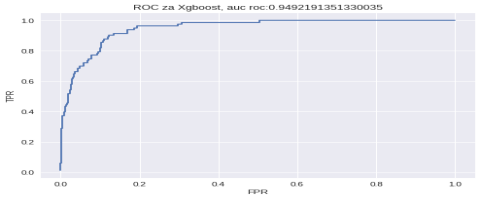
Slika 1.



Slika 2.



Slika 3.



Slika 4.

Iz gornjih ROC krivulja donjeli bi isti zaključako kao iz tablice I.

Za testiranje modela LR, DT, RF i XGB koristili smo *scikit-learn* biblioteku u *Pythonu* dok smo za za XGBSZ implementirali svoju metodu uz korištenje *xgboost* biblioteke također u *Pythonu*.

D. Interpretacija

Nakon prilagodbe modela razmatranjem njegove strukture možemo uvidjeti koje ekonomske mjere bitno sudjeluju pri donošenju predviđanja unutar modela. Za problem predviđanja unaprijed jednu godinu analizirali smo učestalost pojavljivanja određene značajke u stablima ansambla kao indikator te važnosti. Prvih 10 najvažnijih značajki zajedno s odgovarajućim važnostima (postotak od ukupnog broja značajki u šumi) navedeni su u tablici II. Stoga zaključujemo da bi pri analizi poduzeća u svrhu predviđanja propadanja bilo dobro obratiti pozornost na te stavke iz financijskog izvješća.

Značajka	Važnost
Dugotajna imovina	0.122
Obveze za poreze, doprinose i slična davanja	0.045
Kratkotrajna imovina	0.023
Promjena obveza prema zaposlenicima	0.020
Promjena obveza za poreze, doprinose i slična davanja	0.017
Novac u banci i blagajni	0.017
Promjena odgođene porezna	0.016
Postrojenja i oprema	0.014
Novac na početku razdoblja	0.013
Nematerijalna imovina	0.012

Tablica II

IV. Mogući nastavak istraživanja

U budućnosti planiramo poboljšati i proširiti pristup na nekoliko načina opisanih u nastavku.

Kao što smo već spomenili, klasifikacijski zadatak može se redefinirati tako da umjesto binarnog ishoda poduzeće je/nije ušlo u postupak predstečajne nagodbe pokušavamo predvidjeti duljinu trajanja predstečajne nagodbe ili iznos tražbine nakon prijeboja u predstečajnoj nagodbi (kao regresijski ili klasifikacijski problem).

Druga mogućnost proširenja, uz dostupnost potrebnih podataka, je dodavanje dodatnih značajki kao što su npr. neki makroekonomski faktori ili značajke iz mreže vjerovnika o međusobnim dugovanjima među promatranim poduzećima i slično (u dijelu Dosadašnja istraživanja smo vidjeli da to možda može pozitivno utjecati na predikcijsku moć modela).

Što se tiče evaluacije uspješnosti našeg pristupa bilo bi korisno više pažnje posvetiti usporedbi našeg pristupa s ostalim pristupima korištenim u području. Važno je uzeti u obzir da je superiornost nekog predikcijskog modela temeljena na nekoj mjeri uspješnosti na testnom skupu naivan rezultat koji nije dovoljan da bi garantirao da taj model sigurno ima bolje performanse od drugih modela. Za potpunu evaluaciju performansi važno je napraviti

neku vrstu testiranja hipoteza kako bi utvrdili da su uočene razlike u performansama statistički značajne, a ne samo slučajna posljedica ekperimentalnog setupa ([4]). Npr. u [2] koriste Friedmanov test za rangiranje algoritama u području credit-scoringa.

U modelu s kojim pristupamo problemu u ovom projektu pojavljuje se značajan broj hiperparametara pa bi u budućem istraživanju trebalo posvetiti više vremena odabiru modela, tj. optimizaciji hiperparametara. Postoji nekoliko tehnika kojima se može pristupiti ovom problemu. Xia et al. u [9] kao i mi primjenjuju XGBoost metodu, ali za problem credit-scoringa. U svome radu oni su proveli optimizaciju hiperparametara s nekoliko metoda te došli do rezultata da Bayesova optimizacija hiperparametra radi bolje nego metode slučajnog pretraživanja, pretraživanja rešetke i ručnog pretraživanja.

Također, bilo bi dobro posvetiti više vremena interpretaciji modela. Individualna stabla vrlo su interpretabilna no nemaju veliku sposobnost predviđanja dok ansambli imaju puno bolje predikcijske sposobnosti, ali cijena toga je veliki gubitak interpretabilnosti. Postoje načini poboljšanja interpretabilnosti modela stabala. Hara et al. u [5] predlažu metodu postprocesiranja koja poboljšava interpretabilnost ansambla stabala. Metoda se sastoji od toga da se nakon što je naučen kompleksan model ansambla on aprokisimira jednostavnijim modelom koji je interpretabilan.

LITERATURA

[1] E. Alfaro, N. Garcia, M. Gamez and D. Elizondo (2008), "Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks," Decision Support Systems, vol. 45 (1), pp. 110–122.

[2] J. Abellan and J. G. Castellano, "A comparative study on base classifiers in ensemble methods for credit scoring," Expert Systems with Applications, vol. 73, pp. 1–10

[3] T. Fawcett (2004), "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers," Pattern Recognition Letters, vol. 31 (8), pp. 1–38

[4] V. Garcia, A. I. Marques and J. S. Sanchez (2015), "An insight into the experimental design for credit risk and corporate bankruptcy prediction systems," Journal of Intelligent Information Systems, vol 44. (1), pp. 159–189

[5] S. Hara and K. Hayashi (2016), "Making Tree Ensembles Interpretable"

[6] D. Liang, C. Lub, C. Tsaic and G. Shiha (2016), "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study," European Journal of Operational Research, vol 252. (2), pp. 561–572

[7] L. Nanni and A. Lumini (2009), "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring," Expert Systems with Applications, vol. 36 (2), pp. 3028–3033

[8] M. A. Rujoub, D. M. Cook and L. E. Hay (1995), "Using Cash Flow Ratios To Predict Business Failures," Journal of Managerial Issues, vol. 7 (1), pp. 75–90

[9] Y. Xia, C. Liu, Y. Li and N. Liu (2017), "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," Expert Systems with Applications, vol. 78, pp. 225–241

[10] M. Zieba, S. K. Tomczak and J. M. Tomczak (2016), "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," Expert Systems with Applications, vol. 58, pp. 93–101