Algoritmi strojnog učenja za analizu preživljenja

Vrdoljak, Ivana

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Ivana Vrdoljak

ALGORITMI STROJNOG UČENJA ZA ANALIZU PREŽIVLJENJA

Diplomski rad

Voditelj rada:
dr. sc. Tomislav Šmuc

Zagreb, rujan 2019.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Od srca zahvaljujem roditeljima i sestrama koji su bili uz mene tijekom svakog stresnog ispita i veselili se sa mnom prilikom svake uspješno prijeđene stepenice. Poseban doprinos dale su moje bake koje su sa mnom u mislima polagale svaki ispit. Cijelo studiranje bilo je puno lakše i veselije uz divne kolege i prijatelje koje sam stekla na fakultetu. Hvala svim mojim prijateljima uz koje je svaki moj akademski uspjeh bio još veći. Hvala mentoru i asistentima koji su mi mnogo pomogli u izradi ovog rada.

Sadržaj

Sadržaj **iv**

Uvod **1**

1 Uvod u analizu preživljenja **2**
1.1 Terminologija i oznake 2
1.2 Cenzuriranje 3
1.3 Funkcije doživljenja i rizika 4

2 Statističke metode za analizu preživljenja **7**
2.1 Neparametarske metode 8
2.2 Semi parametarski modeli 15
2.3 Parametarski modeli 20

3 Metode strojnog učenja **25**
3.1 Stablo preživljenja 26
3.2 Ansambli 26
3.3 Metoda potpornih vektora 29
3.4 Mjere uspješnosti modela kod analize preživljenja 30

4 Analiza preživljenja za Kkbox podatke **33**
4.1 Programski alati 35
4.2 Opisna statistika 37
4.3 Modeliranje 43
4.4 Zaključak i otvorena pitanja 47

Bibliografija **48**

Uvod

U ovom diplomskom radu proučava se analiza podataka preživljenja. Jedan od glavnih izazova kod analize takvih podataka je pojava cenzuriranja. Kod nekih primjera događaj od interesa se ne dogodi tijekom razdoblja praćenja. Razvijene su različite statističke metode te metode strojnog učenja kojima se učinkovito rješavaju ovakvi problemi.
U poglavlju 1 opisana je terminologija korištena u ovakvim analizama. Definirani su pojmovi i funkcije na kojima se oslanjaju modeli analize preživljenja.
U poglavlju 2 dan je pregled svih statističkih metoda za analizu preživljenja. Opisane su glavne prednosti i mane parametarskih, neparametarskih i semi-parametarskih metoda.
U poglavlju 3 su opisane metode strojnog učenje prilagođene za cenzurirane podatke.
U poglavlju 4 primijenjene su opisane metode na podacima iz stvarnog svijeta.

Poglavlje 1

Uvod u analizu preživljenja

Analiza preživljenja je skup statističkih procedura za analizu podataka s ciljem analize i modeliranja podatka. Rezultat je vrijeme pojave događaja od interesa. Prilikom analize preživljenja unaprijed je potrebno definirati vrijeme promatranja te događaj od interesa.

1.1 Terminologija i oznake

U uvodu će biti definirani osnovni matematički pojmovi i oznake vezane za analizu preživljenja. Budući da je ovakva analiza najčešće prisutna kod medicinskih istraživanja pojmovi koji se koriste su poprimili medicinsku terminologiju. Bez smanjenja općenitosti subjekt ove analize nazivat ćemo osoba, događaj interesa odnosit će se na smrt , a nedostatak tog događaja preživljenje.
Vrijeme preživljenja, koje označavamo s T je slučajna varijabla koja predstavlja vrijeme koje je osoba preživjela. Budući da T predstavlja vrijeme, varijabla je strogo pozitivna. Vrijeme preživljenja u ovom slučaju odgovara zapisu osobe te označava vremenski period obično od početka promatranja sve do događaja od interesa, često je izraženo u mjesecima ili godinama. Alternativno, vrijeme preživljenja može označavati i dob pojedinca kada se događaj dogodio.
Oznaka t predstavlja specifičnu vrijednost vremena od interesa slučajne varijable T. Slučajna varijabla koja indicira smrt ili cenzuriranje označava se grčkim slovom δ te poprima vrijednosti 0 ili 1 . Odnosno, δ=1 predstavlja smrt, tj. događaj interesa unutar perioda promatranja, a δ=0 cenzurirano vrijeme preživljenja.
Ključan problem koji se pojavljuje u analizi preživljenja je cenzuriranje.

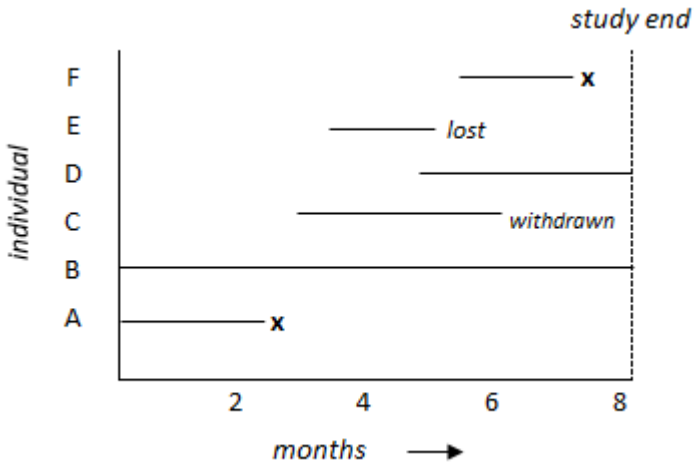
1.2 Cenzuriranje

U longitudinalnoj analizi točno vrijeme preživljenja je poznato samo kada se događaj od interesa dogodi unutar perioda promatranja. Za sve ostale podatke ne znamo što se dogodilo nakon vremena prestanka promatranja, odnosno znamo da se događaj od interesa nije dogodio do kraja promatranja. Takve podatke nazivamo cenzuriranim opažanjima. Pozitivna strana ovakve analize je što takve podatke zadržavamo u analizi i na njima gradimo modele, dok se kod nekih drugih analiza ovakvi podaci u potpunosti zanemare i izbace iz israživanja.
Tri su glavna razloga zašto dolazi do cenzuriranja:

(1) osoba nije doživjela događaj od interesa do kraja perioda promatranja

(2) osoba je „izgubljena" tijekom perioda promatranja

(3) osoba se povukla iz istraživanja



Slika 1.1: Prikaz vremena preživljenja za šest osoba, izvor [7]

Sve tri situacije grafički su prikazane na slici 1.1, pri tome x označava da se dogodio događaj od interesa. Podatak osobe A sadrži informacije od početka promatranja do događaja od interesa koji je bio u trećem mjesecu, taj podatak nije cenzuriran. Podatak osobe B također sadrži informacije od početka promatranja, ali događaj od interesa se nije dogodio te je taj podatak cenzuriran i možemo reći da je vrijeme preživljenja bar osam mjeseci. Dakle, od šest promatranih osoba, kod dvije (osobe A i F) dogodio se događaj od interesa,

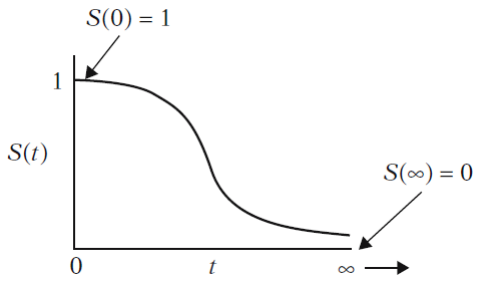
a kod ostale četiri podaci su cenzurirani (osobe B, C, D i E).
Cenzuriranje možemo podijeliti u 3 grupe, ovisno o pojavi cenzuriranja:

1. desno cenzurirani podaci , kada se događaj dogodio nakon kraja perioda promatranja
2. lijevo cenzurirani podaci, kada se događaj dogodio prije početka perioda promatranja
3. intervalno cenzurirani podaci su oni za koje ne znamo točno vrijeme događaja, ali znamo da se dogodio unutar određenog intervala koji promatramo.

Na slici 1.1 vidimo da je na ovom primjeru primjenjeno desno cenzuriranje te iako se druge dvije navedene metode koriste, najviše je korišteno upravo desno cenzuriranje.
Dvije funkcije od posebnog interesa kod ove analize su funkcija doživljenja (engl. survival function) i funkcija rizika (engl. hazard function).

1.3 Funkcije doživljenja i rizika

Definicija 1.3.1. ***Funkcija doživljenja*** je vjerojatnost da je vrijeme doživljenja veće ili jednako nekom vremenu t, tj. predstavlja vjerojatnost da je neka osoba živa u vremenu t. Konvencionalno oznaka za funkciju doživljenja je S: $\mathbb{R} \rightarrow [0,1]$, a njen oblik je
$S(t) = P(T > t)$.



Slika 1.2: Teorijski prikaz funkcije doživljenja

Svojstva funkcije doživljenja su:

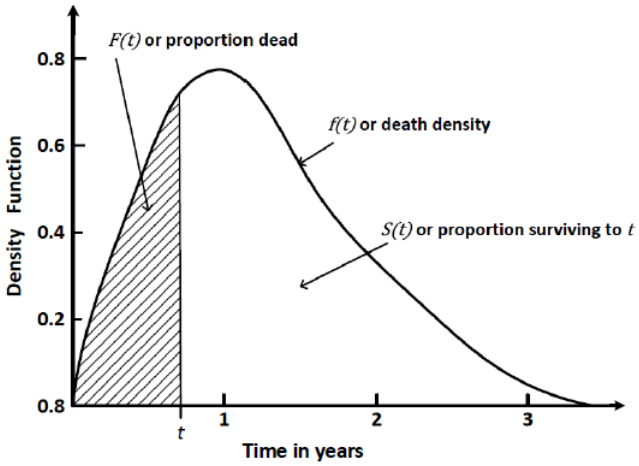
1. funkcija je strogopadajuća
2. Za t=0 vrijedi S(t) = S(0) = 1 , odnosno na početku promatranja sve osobe su žive
3. Za $t = \infty$ vrijedi $S(t) = S(\infty) = 0$, odnosno ako vrijeme ide u beskonačnost vjerojatnost da osoba preživi jednaka je nuli

U praksi graf funkcije doživljenja ima stepenasti oblik, a ne oblik glatke krivulje kao na slici 1.2
.

Definicija 1.3.2. ***Kumulativna funkcija distribucije*** je vjerojatnost da je vrijeme doživljenja manje od nekog vremena t, tj. predstavlja udio populacije koja je umrla do vremena t, zato se naziva još i kumulativnom funkcijom distribucije smrti. Definirana je kao
F(t) = 1 - S(t) = P(T≤t).

Funkcija gustoće smrti definirana je kao:

- f(t) = $\frac{d}{dt}$ F(t) za neprekidne slučajeve
- f(t) = $\frac{F(t+\Delta t)-F(t)}{\Delta t}$, gdje Δt predstavlja mali vremenski interval kod diskretnih slučajeva



Slika 1.3: Prikaz odnosa funkcija za analizu preživljenja, izvor: [7]

Definicija 1.3.3. ***Funkcija rizika*** , znana kao i intenzitet smtnosti te kondicionalna stopa neuspjeha je stopa događaja od interesa u vremenu t pod pretpostavkom da se još do tad nije dogodio. Funkciju rizika definiramo kao h: $\mathbb{R} \to [0,\infty\rangle$,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

Funkcija je neograničena odozgo :

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)} \geq 0$$

Vrijedi:

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}(1 - S(t)) = -\frac{d}{dt}S(t)$$

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}S(t) \cdot \frac{1}{S(t)} = -\frac{d}{dt}(\ln S(t))$$

$$\ln S(t)|_0^u = -\int_0^u h(t)dt$$

$$\ln S(u) - \ln S(0) = -\int_0^u h(t)dt$$

pri čemu je

$$S(0) = P(T \geq 0) = 1$$

pa vrijedi

$$\ln S(u) = -\int_0^u h(t)dt$$

$\Rightarrow S(t) = e^{-H(t)}$, gdje H(t) označava kumulativnu funkciju hazarda

$$H(t) = \int_0^t h(u)du$$

U literaturi se katkad funkcija $-\ln S(\cdot)$ naziva funkcijom rizika.
Funkcija rizika u svakoj točki t direktno odgovara intuitivnom shvaćanju rizika, tj. da se neki događaj dogodio baš u vremenu t.
Za razliku od grafa funkcije doživljenja, graf funkcije rizika ne mora početi u 1 i ne teži k 0 nego za različite vrijednosti t može padati ili rasti. Kao što smo već pokazali, postiže nenegativne vrijednosti i nije ograničena odozgo. Sve su tri funkcije ekvivalentni način zadavanja i dovoljno je imati zadanu jednu.
Sažetak veza:

(1) f(t) = $-\frac{d}{dt}S(t)$

(2) h(t) = $\frac{f(t)}{S(t)}$

(3) S(t) = $e^{-H(t)} = e^{-\int_0^t h(u)du}$

Opisne mjere za analizu preživljenja

Opisne mjere koje nam mogu pomoći na početku analize podataka je prosjek hazarda *(eng. average hazard rate)* te prosjek vremena doživljenja *(eng. average survival time)* .
Prosjek hazarda dan je izrazom $\bar{h} = \frac{\text{broj smrti}}{\sum_{i=1}^n t_i}$, a prosjek vremena doživljenja s $\bar{T} = \frac{\sum_{i=1}^n t_i}{n}$, pri čemu u sumaciju ne ulaze cenzurirane obzervacije.

Poglavlje 2

Statističke metode za analizu preživljenja

Obzirom na odnos između broja primjera za učenje i broja parametara modela, metode možemo podijeliti na parametarske, semi-parametarske i neparametarske. Na slici 2.1 možemo vidjeti sažetak svih statističkih metoda te njihove prednosti i mane.

TIP	PREDNOSTI	MANE	METODE
Neparametarske	Učinkovitije kada ne postoji pripadna teoretska distribucija	Teško interpretabilne, ne daju pouzdane procjenitelje	Kaplan - Meier Nelson - Aalen Životne tablice
Semi - parametarske	Nije potrebno znanje o distribuciji podataka	Nepoznata distribucija izlazne varijable, nekad teško interpretabilno. Moraju vrijediti određene pretpostavke nad podacima.	Cox model Regularizirani Cox Vremenski zavisan Cox
Parametarske	Lako ih je interpretirati, učinkovitije i točnije kada vremena preživljenja prate određenu distribuciju.	Kada je pretpostavka o distribuciji netočna, nekonzistentne su i ne daju optimalan rezultat	Tobit Buckley - James Penalizirana regresija AFT

Slika 2.1: Sažetak statističkih metoda, izvor:[10]

2.1 Neparametarske metode

Neparametarske metode korištene su kada teorijske distribucije ne opisuju dovoljno dobro podatke ili kad pretpostavka o proporcionalnosti hazarda nije zadovoljena. Neparametarske metode su : Kaplan-Meier (KM), Nelson - Aalen te životne tablice.

Metoda životnih tablica za procjenu funkcije doživljenja koristi se za veće skupove podataka čije vrijeme događaja nije precizno određeno i vrijeme možemo podijeliti u nekoliko vremenskih intervala dok je Kaplan-Meier metoda pogodnija za male skupove podataka koji precizno mjere vrijeme događaja. Nadalje, Nelson-Aalen metoda koristi se za analizu kumulativnog hazarda dok je za analizu funkcije doživljenja preferirana Kaplan -Meier metoda.

Kaplan - Meier metoda

Za analizu preživljenja funkcija doživljenja i pripadajući graf su najkorišteniji. 1958. godine Kaplan i Meier napravili su Kaplan - Meier krivulju i *product - limit* (PL) procjenitelj za procjenu funkcije doživljenja koristeći stvarnu duljinu promatranog vremena.
Pretpostavimo da se događaj od interesa može pojaviti u k, $k \in \mathbb{N}$ različitih trenutaka, neka su $t_1 < t_2, ... < t_k$ vremena promatranja za svih n (k ≤ n) primjera. Postoje također i cenzurirane primjeri, čiju sumu za svaki interval t_i $i = 1, ..., k$ označavamo s q_i, a sumu događaja od interesa (smrt) s m_i. Rizični skup, $R(t_{(i)})$, je skup onih kojima je vrijeme preživljenja veće ili jednako $t_{(i)}$. Odnosno, izračunava se na sljedeći način: $R(t_{(i)}) = R(t_{(i-1)}) - q_{i-1} - m_{i-1}$.
Tablica koju koristimo za Kaplan - Meier metodu je oblika:

vremena doživljenja	#događaja od interesa	#cenzuriranih unutar $(t_{(j)}, t_{(j+1)})$	skup rizičnih
$t_{(j)}$	m_j	q_j	$R(t_{(j)})$
$t_{(0)}$=0	m_0=0	q_0	$R(t_{(0)})$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
...	...	...	...
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$

Tablica 2.1: Tablični (sortirani) prikaz podataka

Broj rizičnih obzervacija prije vremena t_j označavamo s n_j. Vjerojatnost preživljenja vremena t_j dano je izrazom $P(T_j) = \frac{n_j - m_j}{n_j}$.

Neka je $t \in \mathbb{R}$ realan broj za koji vrijedi $t_j \leq t \leq t_{j+1}$, $j \in \{1, 2, ..., k-1\}$. Pretpostavimo

da su svi događaji od interesa nezavisni, tada vrijedi:

$$\begin{aligned}
S(t) &= P(T > t) = P(\{T > t\} \cap \{T > t_{(j)}\}) = P(T > t | T > t_{(j)}) \cdot P(T > t_{(j)}) \\
&= P(T > t | T > t_{(j)}) \cdot P(\{T > t_{(j)}\} \cap \{T > t_{(j-1)}\}) \\
&= P(T > t | T > t_{(j)}) \cdot P((T > t_{(j)} | T > t_{(j-1)}) \cdot P(T > t_{(j-1)}) = ... \\
&= P(T > t | T > t_{(j)}) \cdot P((T > t_{(j)} | T > t_{(j-1)}) ... P((T > t_{(2)} | T > t_{(1)}) \cdot P(T > t_{(0)}) \\
&= \prod_{i=1}^{j} P(T > t_{(i)} | T \geq t_{(i)})
\end{aligned}$$

Kaplan Meier procjenitelj funkcije je:

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - m_j}{n_j}$$

Naime, pod pretpostavkom da se događaji od interesa događaju točno u vremenima t_j, $j \in \{1, 2, ..., k-1\}$, promotrimo interval $\langle t_{(j)} - \Delta t, t_{(j)}]$. Vjerojatnost događaja od interesa unutar intervala pod pretpostavkom preživljenja do trenutka $t_{(j)} - \Delta t$ je $\frac{m_j}{n_j}$. Iz toga slijedi da je vjerojatnost preživljenja $\frac{n_j - m_j}{n_j}$. Promotrimo sad interval $\langle t_{(j)}, t_{(j+1)} - \Delta t \rangle$, u tom intervalu znamo da se nije dogodio niti jedan događaj od interesa.
Zbog toga slijedi: $P(T > t_{(j)} - \Delta t | T \geq t_{(j+1)} - \Delta t) = \frac{n_j - m_j}{n_j}$ $(\Delta t \to 0)$.

Kaplan Meier procjenitelj funkcije doživljenja zovemo i *product - limit* aproksimacija i možemo ga zapisati na sljedeći način:

$$\hat{S}(t_{(j)}) = \hat{S}(t_{(j-1)}) \cdot P(T > t_{(j)} | T \geq t_{(j)})$$

Ekvivalencija tih dviju formula se može dokazati koristeći svojstvo uvjetne vjerojatnosti.

Primjer 2.1.1. *U tablici su dani podaci izmjereni tijekom medicinskog promatranja pacijenata oboljelih od leukemije. Pacijenti su podijeljeni u dvije grupe, oni koji su primili placebo i oni koji su primili lijek. Pomoću Kaplen - Meier metode usporedimo funkcije doživljenja te dvije grupe. (primjer je preuzet iz [7])*

grupa 1 (n=21) - placebo	grupa 2 (n=21) - lijek
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Tablica 2.2: Vremena remisije pacijenata

Znakom + označeni su cenzurirani podaci. Prvo, izračunajmo spomenute dvije opisne mjere:

$$\bar{T}_1 = 17.1, \bar{T}_2 = 8.6, \bar{h}_1 = 0.025, \bar{h}_2 = 0.115$$

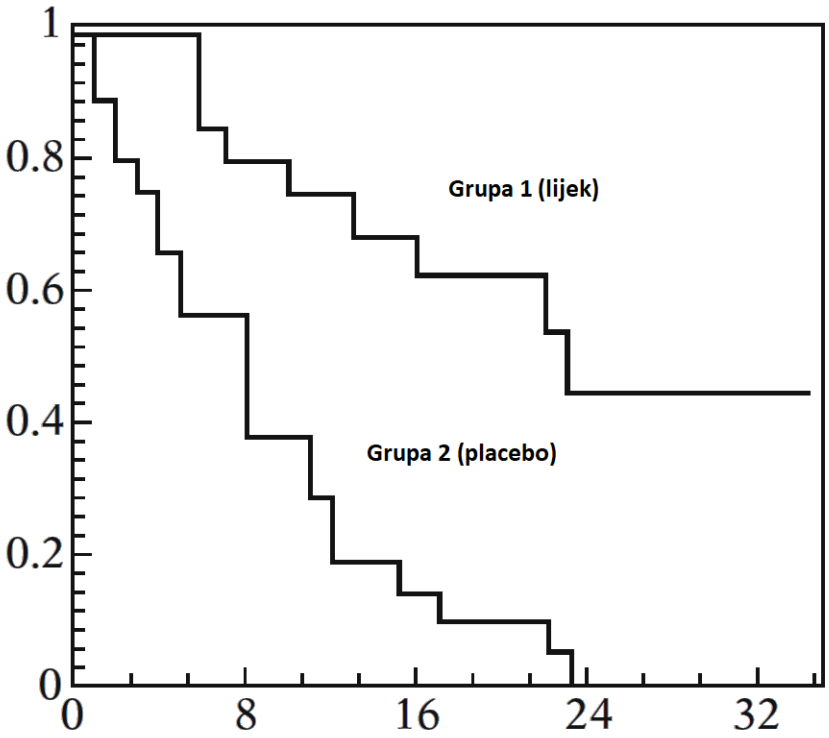
Na temelju ovih podataka vidimo kako pacijenti u grupi 1 imaju veću vjerojatnost preživljenja, no treba uzeti u obzir da su ovo samo opisne mjere koje ne uzimaju u obzir različita vremena nego sumarno daju pregled nad podacima. Ispunjene tablice koju koristimo za Kaplan - Meier metodu za obje grupe je:

$t_{(j)}$	n_j	m_j	q_j	$\hat{S}(t_{(j)})$
0	21	0	0	1
6	21	3	1	0.8571
7	17	1	1	0.8067
10	15	1	2	0.7529
13	12	1	0	0.6902
16	11	1	3	0.6275
22	7	1	0	0.5378
23	6	1	5	0.4482

Tablica 2.3: KM tablica - grupa 1

$t_{(j)}$	n_j	m_j	q_j	$\hat{S}(t_{(j)})$
0	21	0	0	1
1	21	2	0	0.90
2	19	2	0	0.81
3	17	1	0	0.76
4	16	2	0	0.67
5	14	2	0	0.57
8	12	4	0	0.38
11	8	2	0	0.29
12	6	2	0	0.19
15	4	1	0	0.14
17	3	1	0	0.10
22	2	1	0	0.05
23	1	1	0	0.00

Tablica 2.4: KM tablica - grupa 2



Slika 2.2: KM krivulje, izvor: [7]

Graf Kaplan -Meier procjenitelja funkcije doživljenja je stepenasta funkcija kojoj su procjenjene vjerojatnosti preživljenja konstantne između dva susjedna vremena i padajuće su u odnosu na vrijeme. Obje KM krivulje za grupe 1 i 2 prikazane su na istom grafu 4.1 radi lakše usporedbe . Može se vidjeti kako je KM krivulja grupe 1 stalno iznad KM krivulje grupe 2. Ovaj prikaz nam sugerira kako je prognoza za preživljenje pacijenta grupe 1 puno povoljnija nego pacijenta koji prima placebo. Štoviše, kako se broj tjedana povećava, udaljenost između krivulja je sve veća i veća, što implicira da što tjedni remisije prolaze, sve veći efekt na preživljenje ima lijek u odnosu na placebo.

Nelson - Aalen procjenitelj

Nelson - Aalen procjenitelj je metoda procjene funkcije kumulativnog hazarda i dana je s izrazom:

$$\hat{H(t)} = \sum_{j=1}^{k} \frac{m_j}{n_j}$$

Dakle, Nelson-Aalen procjenitelj je rastuća, neprekidna, stepenasta funkcija sa skokovima $\frac{m_j}{n_j}$ između dva susjedna vremena. Iz toga slijedi da je procjenitelj funkcije doživljenja oblika

$$\hat{S}(t) = \prod_{j=1}^{k} \exp(-d_j/n_j)$$

Budući da vrijedi

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + ...$$

te je aproksimacija za dovoljno male vrijednosti x jednaka 1-x. Iz tog slijedi da je $\exp(-m_j/n_j) \approx 1 - (m_j/n_j) = (n_j - m_j)/n_j$, dok god su vrijednosti od m_j dovoljno male u odnosu na n_j, što je istina, osim u zadnjim vremenima preživljenja. Dakle, možemo reći da je tada Kaplan - Meier procjenitelj aproksimacija Nelson - Aalen procjenitelja. Štoviše, znamo i da je Neelson-Aalen procjenitelj uvijek veći od Kaplan-Meier procjenitelja u bilo kojem vremenu promatranja zato što vrijedi $e^{-x} \geq 1 - x$, za sve vrijednosti x.

Neparametarske metode koristimo kada želimo usporediti vrijeme preživljenja za više grupa te pomoću njih dobijemo početni uvid u svojstva podataka. No, budući da u obzir uzimaju samo varijablu vrijeme, a ostale varijable ignoriraju, ne možemo ih koristiti za predikciju.

Log-Rank test

Kako bismo procijenili jesu li KM krivulje, tj. funkcije doživljenja za dvije ili više grupa statistički jednake koristimo metode usporedbe od kojih je najpopularnija Log - Rank test. Log Rank test temelji se na usporebi promatranih i očekivanih vrijednosti, a hipoteza je oblika:

$$H_0 : S_1(t) = S_2(t) = ... = S_G(t)$$

$$H_1 : \text{bar jedna se razlikuje}$$

, pri čemu G (≥ 2) označava broj grupa.
Bez smanjenja općenitosti u nastavku je prikazana metoda testiranja za dvije grupe (G=2). Očekivano vrijeme za grupu 1 (e_{1j}) je omjer ukupnog broja obzervacija unutar obje grupe koji su u riziku u vremenu t_j tj. $n_{1j}/(n_{1j} + n_{2j})$ multipliciran s ukupnim brojem događaja od interesa u tom trenutku u obje grupe ($m_{1j} + m_{2j}$).

$$e_{1j} = (\frac{n_{1j}}{n_{1j} + n_{2j}}) \times (m_{1j} + m_{2j})$$

$$e_{2j} = (\frac{n_{2j}}{n_{1j} + n_{2j}}) \times (m_{1j} + m_{2j})$$

Log- Rank statistika za dvije grupe je:

$$\frac{(O_i - E_i)^2}{\hat{Var}(O_i - E_i)}, \text{ i = 1,2}$$

Izraz za procjenu varijance sadrži broj rizičnih obzervacija (n_{ij}) i broj događaja od interesa (m_{ij}) u vremenu t_j. Sumacija je po svim različitim vremenima t.

$$\hat{Var}(O_i - E_i) = \sum_j \frac{n_{1j}n_{2j}(m_{1j} + m_{2j})(n_{1j} + n_{2j} - m_{1j} - m_{2j})}{(n_{1j} + n_{2j})^2((n_{1j} + n_{2j} - 1)}$$

Aproksimativna formula koja se može koristiti kada izračun nije programski, a dana je s:

$$X^2 = \sum_{i=1}^{\text{G}} \frac{(O_i - E_i)^2}{E_i}$$

Općenito, log-rank statistika je aproksimativno χ^2 s G-1 stupnjem slobode te se P-vrijednost čita iz tablice χ^2 distribucije. Log-rank test daje jednake težine svim obzervacijama, ali također je važno napomenuti da je ovaj test najprikladniji kada je ispunjena pretpostavka o proporcionalnom hazardu.

Peto test

Za razliku od log-rank testa, Peto test nema pretpostavku o proporcionalnosti hazarda. Peto test statistika je modifikacija log-rank statistike i koristi procjenitelj funkcije doživljenja kao težinu.

$$\textit{Peto test statistika} = \frac{(\sum_j \hat{S}(t_j)(m_{ij} - m_j\frac{n_{ij}}{n_j}))^2}{\sum_{j=1}^{k} \hat{S}(t_j)^2 \frac{n_{1j}n_{2j}m_j(n_j - m_j)}{n_j^2(n_j - 1)}}$$

Procjenitelj funkcije doživljenja jednak je KM procjenitelju funkcije preživljenja pomnoženim s $\frac{n_j}{n_j+1}$.

$$\hat{S}(t) = \prod_{t_j<t}(1 - \frac{m_j}{n_j + 1})$$

Za veliki uzorak Peto test statistika također teži χ^2 distribuciji s G-1 stupnjem slobode.
Peto test stavlja naglasak na početno vrijeme promatranja i na te obzervacije stavlja veću težinu dok je kod log-rank testa naglasak na kraju promatranja. Ovisno o tome koji dio promatranja želimo naglasiti koristimo jedan od ova dva testa. Prednost log-rank testa je što je precizniji kada cenzurirani podaci nisu iz iste distribucije za sve grupe. Još neki od testova za usporedbu funkcija doživljenja više grupa, s težinom w_j su:

(1) Gehan (generalizirani) Wilcoxon test, $w(t_j) = n_j$

(2) Tarone-Ware test, $w(t_j) = \sqrt{n_j}$

(3) Fleming-Harrington test, $w(t_j) = \hat{S}(t_{j-1})^p(1 - \hat{S}(t_{j-1})^q$, pri čemu je $\hat{S}(t)$ KM procjenitelj

Od nabrojenih metoda najviše fleksibilnosti kod pripadnih težina daje Fleming-Harrington test jer korisnik određuje vrijednost parametara p i q.

Pri izboru metode za usporedbu treba obratiti pažnju na proporcionalnost hazarda, proporcije cenzuriranih podataka, veličinu uzorka te distribuciju podataka. Izbor metode trebao bi biti apriori, bez mogućnosti manipulacije podacima.

2.2 Semi parametarski modeli

U kategoriji semi parametarskih modela najkorišteniji pristup za analizu regresije podataka preživljenja je Cox model. On se značajno razlikuje od drugih modela jer koristi pretpostavku o proporcionalnosti hazarda. 1972. godine je u znanstvenom članku naziva *„Regression models and life tables"* znanstvenik David Cox prezentirao model proporcionalnog hazarda koji kaže da je uvjetni hazard obzirom na vrijeme preživljenja i zadanim kovarijatama produkt osnovne funkcije hazarda i eksponencijalne regresijske funkcije kovarijata.

Cox model

Kao hibrid parametarskih i neparametarskih metoda, Cox model može sadržavati konzistentnije procjenitelje pod više pretpostavka u usporedbi s parametarskim modelima te imati preciznije procjenitelje u usporedbi s neparametarskim. Za razliku od parametarskih modela, znanje o distribuciji događaja od interesa nije nužno, ali se pretpostavlja da varijable imaju eksponencijalni učinak na ishod. Cox model se naziva semi parametarski model jer je umnožak dvije funkcije, od kojih je jedna osnovni (*eng. baseline*) hazard, a druga se sastoji od linarne sume vektora koeficjenata β i varijabla.

Definicija 2.2.1. Neka je $\mathbf{X}=(X_1, X_2, ..., X_n)$ vektor nezavisnih varijabli koje odgovaraju pojedincu s odgovarajućim opažanjima $X_i=(x_{i1}, ..., x_{ip})$ te $\beta=(\beta_1, \beta_2, ..., \beta_p)$ vektor koeficijenata, tada je osnovni Cox model

$$h(t, X_i) = h_0(t)e^{X_i\beta} \tag{2.1}$$

Funkcija $h_0(t)$ naziva se osnovnim (*eng. baseline*) hazardom.

Ako su sve varijable X_i jednake 0, tada je model jednak osnovnom hazardu, otkuda i naziv osnovni (početni) hazard. Budući da $h_0(t)$ nije specificirana i ovisi o vremenu t, Cox model nije izražen samo preko nepoznatih parametara poput parametarskih modela.
Pretpostavka kod osnovnog Cox modela je da su varijable vremenski nezavisne. Ako razmatramo podatke koji su vremenski zavisni, pretpostavka o proporcionalnom hazardu nije zadovoljena, ali možemo koristiti takozvani prošireni Cox model. Za varijable kažemo da su vremenski nezavisne ako se ne mijenjaju tijekom vremena. Primjeri takvih varijabli su spol, krvna grupa. Postoje dvije vrste vremenski zavisnih varijabli: interne te eksterne. Eksterne varijable se ne mijenjaju često i ostvarenje događaja od interesa ne ovisi o promjeni varijable, to su na primjer dob, težina. Takve varijable, budući da se ne mijenjaju često možemo uzeti u model. Interne varijable su one koje mogu biti mjerene do događaja od interesa (tlak, otkucaj srca...).

Glavni razlog popularnosti Cox modela leži u tome što, iako funkcija osnovnog hazarda nije specificirana, može pronaći izuzetno dobre procjenitelje koeficjenata, relativnog hazarda te funkcije doživljenja za puno različitih skupova podataka. Cox model je „siguran" odabir kada ne znamo dolaze li podaci iz neke konkretne parametarske razdiobe jer zbog svoje robusnosti jako dobro aproksimira parametarske modele.

Kako bismo mogli koristiti Cox model, nužno je da vrijedi pretpostavka o proporcionalnosti hazarda. Proporcionalnost hazarda znači da je relativni hazard za dvije grupe (grupe pacijenata koji uzimaju lijek i onima kojima je dan placebo) X_1 i X_2 konstantan kroz vrijeme:

$$\hat{HR} = \frac{h(t, X_1)}{h(t, X_2)} = \frac{h_0(t)e^{X_1\beta_i}}{h_0(t)e^{X_2\beta_i}} = exp[(X_1 - X_2)\beta_i]$$

Hazard za svaku osobu je fiksna proporcija hazarda od bilo koje druge osobe i konstantna je u vremenu. Hazard se može mijenjati kroz vrijeme, npr. oba mogu rasti povećanjem vremena, ali omjer ostaje konstantan. Prisjetimo se, što je veći hazard to je veća vjerojatnost da će se događaj od interesa dogoditi.
Za binomnu varijablu, uzmimo primjer varijablu pacijent prima lijek , odnosno pacijent prima placebo, relativni hazard je oblika:

$$\exp(\hat{\beta}) = \frac{\text{hazard pacijenata koji su primali lijek u vremenu } t}{\text{hazard pacijenata koji su primali placebo u vremenu } t}$$

, a općeniti oblik za binomnu varijablu je

$$\hat{HR} = exp[\hat{\beta} + \sum \delta_j W_j]$$

pri čemu su $\hat{\beta}$ procjenjeni koeficjenti varijable, a δ_j koeficjenti uz interakciju s drugom varijablom, ako promatramo interakcije. Odnosno proporcionalni hazard znači da se hazard može mijenjati tijekom vremena, ali omjer je konstantan. Kasnije će biti opisano kako provjeriti tu pretpostavku.

Funkcija doživljenja za Cox model dana je izrazom :

$$S(t) = exp(-H_0(t)exp(X\beta)) = S_0(t)^{exp(X\beta)}$$

pri čemu je H_0(t) kumulativna funkcija osnovnog hazarda, a S_0(t)=exp(-H_0(t)) predstavlja osnovnu funkciju doživljenja.

Uzmimo za primjer medicinsko istraživanje, ako je cilj u određenom trenutku otkriti tko je deset najugroženijih pacijenata, jer postoje resursi za zadržati u bolnici samo desetero ljudi, koristit ćemo funkciju hazarda. No ako nas zanima vjerojatnost da se pacijent vrati za 30 dana u bolnicu koristimo osnovnu funkciju hazarda. Zbog toga postoje i različite

metode procjene. Najpopularniju procjenu kumulativne funkcije osnovnog hazarda H_0(t), predložio je Breslow (1972.) i zato se naziva Breslow procjenitelj:

$$\hat{H}_0(t) = \sum_{t_i<t} \hat{h}_0(t_i)$$

gdje je

$$\hat{h_0}(t_i) = \begin{cases} 1/\sum_{j\in R_i} e^{X_j\beta}, & \text{za } t_i \text{ vrijeme preživljenja} \\ 0, & \text{inače.} \end{cases}$$

R_i pri tome predstavlja skup rizičnih pacijenata u vremenu t_i.

Procjena metodom maksimalne vjerodostojnosti

Jedna od glavih prednosti Cox modela je što možemo procijeniti parametre β bez poznavanja funkcije osnovnog hazarda $h_0(t)$.

Koeficijenti β_i se procjenjuju metodom maksimalne vjerodostojnosti (*engl. maximum likelihood*) i označujemo ih s $\hat{\beta}_i$.

Zbog osnovne funkcije hazarda koja nije specificirana nije moguće koristiti standardnu funkciju vjerodostnojnosti. Cox je predložio funkciju parcijalne vjerodostojnosti oblika

$$L(\beta) = \prod_{j=1}^{N} \left[\frac{exp(X_j\beta)}{\sum_{i\in R_j} exp(X_i\beta)} \right]^{\delta_j} \tag{2.2}$$

Uočimo da kada je δ_j=1 j-ti faktor je uvjetna vjerojatnost dok je za δ_j=0 faktor 1 što znači da ne doprinosi produktu, odnosno cenzurirani događaji ne utječu direktno na vjerodostojnost. Ipak, u rizik setu su sumirani i cenzurirani zapisi pa doprinose produktu u tom dijelu. Zbog lakšeg računanja funkcija vjerodostojnosti se logaritimira:

$$logL(\beta) = \sum_{j=1}^{N} \delta_j \left\{ X_j\beta - log\left[\sum_{i\in R_j} exp(X_i\beta) \right] \right\} \tag{2.3}$$

Maksimalni parcijalni procjenitelj vjerodostojnosti (*eng. maximum partial likelihood estimator (MPLE)*) se uz numeričku Newton-Raphson metodu koristi kako bi iteracijama pronašli procjenitelj koeficjenata $\hat{\beta}$ maksimiziranjem jednadžbe (2.3).

Regularizirani Cox modeli

Razvitkom baza podataka i razvitkom mogućnosti skladištenja podatka, podaci za analize postali su visokih dimenzija. U nekim slučajevima, broj varijabli (p) je skoro pa jednak,

a ponekad i veći, od broja obzervacija (n). Zbog velikog broja varijabli kreiranje modela predikcije je komplicirano jer vrlo lako može doći do overfittinga. To predstavlja opasnost da će model dobro raditi samo na podacima na kojima je napravljen („istreniran"). Pod pretpostavkom da nisu sve varijable značajne za model koriste se razne metode kako bi se smanjio broj varijabli za modeliranje. Efekta zvan sparcity, odnosno rijetke matrice, znači da od koeficjenata svih varijabla koje opisuju model, samo mali broj njih nije jednak nula (tzv. modeli s težinama pritegnutima na nulu). Postoji puno metoda kojima se sparcity podataka postiže, a jedna od metoda je regularizacija funkcije pogreške.
Regularizacijska funkcija l -lorma $l_\gamma : \mathbb{R}^\gamma \to \mathbb{R}$ je oblika

$$l_\gamma(\beta) = ||\beta||_\gamma = (\sum_{i=1}^{P} ||\beta_i||^\gamma)^{\frac{1}{\gamma}}, \gamma > 0$$

Što je manja vrijednost γ , to je rješenje sklonije obliku rijetke matrice, no kada je $0 \leq \gamma < 1$, regularizacijska funkcija nije konveksna te je rješenje optimizacije teže rješivo.

METODA	REGULARIZACIJSKI IZRAZ	PRIMJENA
LASSO	$\sum_{k=1}^{p} \|\beta_k\|$	sparsity
RIDGE	$\sum_{k=1}^{p} \beta_k^2$	korelacija
ELASTIČNA MREŽA (EN)	$\mu \sum_{k=1}^{p} \|\beta_k\| + (1-\mu) \sum_{k=1}^{p} \beta_k^2$	sparsity i korelacija
OSCAR	$\lambda_1 \|\beta\|_1 + \lambda_2 \|T\beta\|_1$	sparsity i graf korelacije

Tablica 2.5: Različiti regularizacijski izrazi korišteni za Cox model

Pomoću regularizatora prikazanih u tablici 2.5 parametar λ može biti odabran tako da utječe na određeno svojstvo podataka. Performanse takvih regulariziranih oblika značajno ovise o λ te optimalni λ_{opt} može se dobiti pomoću unakrsne-validacije.

Provjera proporcionalnosti hazarda

Budući da je svojstvo proporcionalnosti hazarda (skraćeno PH) iznimno bitno kod upotrebe Cox modela opisano je kako provjeriti zadovoljavaju li podaci to svojstvo. PH pretpostavka nije zadovoljena sijeku li se funkcije hazarda 2 ili više kategorija. No, moguće je da se funkcije ne sijeku, a PH pretpostavka nije zadovoljena. Tri su glavna načina kako provjeriti pretpostavku: grafičkim prikazom, korištenjem Z statistike za veliki uzorak (goodnes of fit) te koristeći vremenski zavisne kovarijate. Dva grafička pristupa za provjeru pretpostavke su usporedba log-log funkcija doživljenja te usporedbom opaženih i očekivanih

krivulja preživljenja. Log-log funkcija preživljenja je jednostavno transformacija procjenjene funkcije preživljenja logaritmiranjem dva puta.

$$\text{log-log}\hat{S} = -ln(-ln\hat{S}) = -ln(\int_0^t h(u)du)$$

$$-ln[-lnS(t,X)] = -\sum_{i=1}^{p}\beta_i X_i - ln[-lnS_0(t)]$$

Ako je PH pretpostavka zadovoljena log-log funkcije doživljenja za opažanja su približno paralelne. Ovaj pristup koristan je kad postoji malo kovarijati za koje trebamo provjeriti pretpostavku i za njih možemo koristiti mali broj kategorija. Usporedba opaženih i očekivanih krivulja grafički je analogon goodnes of fit testa. Goodness of fit („dobrota pristajanja") koristi χ^2 test s 1 stupnjem slobode baziran na opaženim i očekivanim vjerojatnostima doživljenja. Proporcionalnost hazarda može se testirati koristeći i vremenski zavisne kovarijate pri čemu je nulta hipoteza da su svi koeficjenti jednaki nula, a testna statistika χ^2 omjera vjerodostojnosti s p stupnjeva slobode, gdje je p broj kovarijata za koje testiramo pretpostavku. Ako nulta hipoteza nije odbačena na temelju p vrijednosti, možemo zaključiti da jedna od kovarijata ne ispunjava PH pretpostavku.

Vremenski zavisni Cox model

Cox model je prilagođen upotrebi kada su u modelu i vremenski zavisne varijable (kod kojih PH pretpostavka nije ispunjena). Neka je dan problem analize preživljenja s P_1 vremenski zavisnih i P_2 vremenski nezavisnih kovarijata označenih s vektorom $(X_1(t), X_2(t), ..., X_{p1}(t), X_1, X_2, ..., X_{p2})$. Cox model s vremenski zavisnim kovarijatama je oblika:

$$h(t,X(t)) = h_0(t)exp\Bigg[\underbrace{\sum_{j=1}^{P_1}\delta_j X_j(t)}_{\text{vremenski zavisan dio}} + \underbrace{\sum_{i=1}^{P_2}\beta_i X_i}_{\text{vremenski nezavisan dio}}\Bigg].$$

Relativni hazard za dvije obzervacije $X_1(t) = (X_{11}(t), X_{12}(t), ..., X_{1p1}(t), X_{11}, X_{12}, ..., X_{1p2})$ i $X_2(t) = (X_{21}(t), X_{1+22}(t), ..., X_{2p1}(t), X_{21}, X_{22}, ..., X_{2p2})$ je izražen s

$$\hat{HR}(t) = \frac{\hat{h}(t,X_2(t))}{\hat{h}(t,X_2(t))} = exp\Bigg[\sum_{j=1}^{P_1}\alpha_j[X_{2j}(t) - X_{1j}(t)] + \sum_{i=1}^{P_2}\beta_i[X_{2i}(t) - X_{1i}(t)]\Bigg].$$

Budući da je prva suma vremenski zavisna, relativni hazard je isto ovisan o vremenu t. Funkcija vjerodostojnosti konstruirana je kao i za osnovni Cox model.

2.3 Parametarski modeli

Kod parametarskih modela složenost modela odosno broj parametara ne ovisi o broju primjera za učenje. Pretpostavka je da vremena doživljenja prate neku određenu teorijsku razdiobu (npr. Gaussovu razdiobu). Učenje se svodi na nalaženje parametara pretpostavljene distribucije, pri čemu broj parametara ne ovisi o broju primjera. Korištenjem parametarskih metoda predviđanje vremena dožviljenja je jednostavno i efikasno. Ukoliko ne postoji odgovarajuća teorijska distribucija za podatke, efikasnije je koristiti neparametarske metode. Najkorištenije funkcije distribucije su eksponencijalna, Weibull i log - logistička distribucija čije su osnovne funkcije koje su potrebne za analizu preživljenja u tablici 2.6 .

DISTRIBUCIJA	S(t)	h(t)	f(t)
Eksponencijalna	$\exp(-\lambda t)$	λ	$\lambda \exp(-\lambda t)$
Weibull	$\exp(-\lambda t^p)$	$\lambda p t^{p-1}$	$\exp(-\lambda t^p)\lambda p t^{p-1}$
Logistička	$\frac{e^{-(t-\mu)/\sigma}}{1+e^{-(t-\mu)/\sigma}}$	$\frac{1}{1+e^{-(t-\mu)/\sigma}}$	$\frac{e^{-(t-\mu)/\sigma}}{(1+e^{-(t-\mu)/\sigma})^2}$
Log - logistička	$\frac{1}{1+\lambda t^p}$	$\frac{\lambda p t^{p-1}}{1+\lambda t^p}$	$\frac{\lambda p t^{p-1}}{(1+\lambda t^p)^2}$

Tablica 2.6: Funkcija doživljenja, funkcija gustoće i funkcija hazarda najkorištenijih distribucija parametarskih modela

Definicija 2.3.1. Slučajna varijabla X ima **eksponencijalnu distribuciju** s parametrom $\lambda > 0$ ako joj je funkcija gustoće

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{x>0.} \\ 0, & \text{inače.} \end{cases} \tag{2.4}$$

Eksponencijalna distribucija najjednostavnija je i jedna od najvažnijih parametarskih metoda budući da je u potpunosti definirana s konstatnim hazardom λ . U ovom slučaju, smrt tj. događaj od interesa je slučajan događaj neovisan o vremenu. Veća vrijednost parametra λ znači veći rizik i kraći period preživljenja. Iz tablice 2.6 se vidi kako je $\ln S(t) = -\lambda t$, tj. kako je veza između logaritma funkcije doživljenja i vremena t linearna. Dakle, kako bi vidjeli prati li vrijeme doživljenja eksponencijalnu funkciju dovoljno je napraviti nacrtati vrijednosti $\ln S(t)$ u ovisnosti o t.

Definicija 2.3.2. Slučajna varijabla X ima **Weibull distribuciju** s parametrima $\lambda > 0$ i $p > 0$, ako joj je funkcija gustoće dana izrazom

$$f(x) = \begin{cases} \lambda p x^{p-1} e^{-\lambda x^p}, & \text{x>0.} \\ 0, & \text{inače.} \end{cases} \tag{2.5}$$

Weibull distribucija najkorištenija je parametarska metoda koja je definirana pomoću dva parametra, $\lambda > 0$ i p >0 . Oblik funkcije hazarda određen je varijablom k, što omogućuje veću fleksibilnost u usporedbi s eksponencijalnim modelom. Za p = 1 funkcija hazarda je konstantna i odgovara eksponencijalnom modelu. Za p < 1, fukcija hazarda je padajuća s obzirom na vrijeme.

Definicija 2.3.3. Slučajna varijabla X ima **logističku distribuciju**, odnosno **log-logističku distribuciju** s parametrima $\mu \in \mathbb{R}$ i $\sigma > 0$, ako joj je funkcija gustoće dana izrazom

$$f(x) = \frac{e^{-(x-\mu)/\sigma}}{(1 + e^{-(x-\mu)/\sigma})^2}$$

za logističku te

$$f(x) = \frac{\lambda p x^{p-1}}{(1 + \lambda x^p)^2}$$

za log-logističku distribuciju.

Logistička i log- logistička distribucija Za razliku od Weibull modela, funkcije hazarda za oba ova modela dopuštaju promjenjivost. Vrijeme doživljenja T i logaritam vremena doživljenja $\ln T$ prate logističku distribuciju. Za k≤0, fukcija hazarda pada povećanjem vremena. Za k>1 funkcija raste do maksimalne vrijednosti te onda pada.

Postoji puno različitih načina za procjenu parametara, neke od metoda su :

(1) metoda maksimalne vjerodostojnosti (MLE)

(2) metoda najmanjih kvadrata

(3) metoda momenata

(4) Cramer Rao metoda

Najkorištenija metoda je metoda maksimalne vjerodostojnosti i slijedi njen opis.

Metoda maksimalne vjerodostojnosti (MLE)

Definicija 2.3.4. Neka je $(x_1, ..., x_n)$ opaženi uzorak za slučajnu varijablu X s gustoćom f(x|β) gdje je $\beta = (\beta_1, \beta_2, ..., \beta_P) \in \Theta \subseteq \mathbb{R}^P$ nepoznati parametar. Definiramo **funkciju vjerodostojnosti** L:$\Theta \longrightarrow \mathbb{R}$ sa

$$L(\beta) := f(x_1|\beta) \cdots f(x_n|\beta), \beta \in \Theta.$$

Vrijednost $\hat{\beta} = \hat{\beta}(x_1, x_2, ..., x_n) \in \Theta$ za koju je

$$L(\hat{\beta}) = \max_{\beta \in \Theta} L(\beta)$$

zovemo **procjena metodom maksimalne vjerodostojnosti**. Statistika $\hat{\beta}(X_1, ..., X_n)$ je **procjenitelj metodom maksimalne vjerodostojnosti** ili kraće MLE.

Pretpostavimo da je broj primjera N s c cenzuriranih i (N-c) necenzuriranih vrijednosti te $\beta = (\beta_1, \beta_2, ..., \beta_P)$ nepoznati parametar. Tada funkcija gustoće smrti f(t) i funkcija doživljenja S(t) mogu biti zapisane kao f($t;\beta$) i S($t;\beta$). Za dan indeks i, ako je podatak cenzuriran, vrijeme doživljenja biti će nepoznato no budući da možemo zaključiti da osoba nije doživjela događaj od interesa prije cenzuriranog vremena C_i, vrijednost funkcije doživljenja S($C_i;\beta$) biti će vjerojatnost blizu 1. U suprotnom, ako se događaj od interesa za indeks i dogodio u vremenu T_i tada će funkcija gustoće smrti $f(T_i;\beta)$ imati visoku vjerojatnost.

Neka je s $\prod_{\delta_i=1} f(t_i;\beta)$ umnožak vjerojatnost svih necenzuriranih primjera i neka je $\prod_{\delta_i=0} S(t_i;\beta)$ umnožak vjerojatnosti cenzuriranih obzervacija. Optimiziranjem funkcije vjerodostojnosti mogu se procijeniti parametri β za svih N primjera:

$$L(t;\beta) = \prod_{i=1}^{n} f(t_i;\beta)^{\delta_i} S(t_i;\beta)^{1-\delta_i}$$

,a budući da je $f(t_i|\beta) = h(t_i;\beta)S(t_i;\beta)$ vrijedi:

$$L(t;\beta) = \prod_{i=1}^{n} h(t_i;\beta)^{\delta_i} S(t_i;\beta).$$

Primjer 2.3.5. *Neka je X1, X2, . . . , Xn slučajni uzorak iz eksponencijalne razdiobe s parametrom β. Korištenjem metode maksimalne vjerodostojnosti procijenimo nepoznati parametar β.*
$X \sim E(\beta) \Rightarrow f(t;\beta) = \beta e^{-\beta t}, S(t;\beta) = e^{-\beta t}$
Neka je x1, x2, . . . , xn opaženi uzorak. Tada vrijedi $x_1, x_2, ..., x_n \in$ N i

$$L(t;\beta) = \prod_{i=1}^{n} \beta^{n-c} e^{-\beta t}$$

Funkcija x $\rightarrow lnx$ je strogo rastuća pa je dovoljno maksimizirati funkciju

$$l(\beta) = \ln L(t;\beta) = (n-c) \cdot \ln\beta - \beta \sum_{i=1}^{n} t_i$$

$$l'(\beta) = \frac{n-c}{\beta} - \sum_{i=1}^{n} t_i$$

$$l'(\beta) = 0 \Leftrightarrow \beta = \frac{n-c}{\sum_{i=1}^{n} t_i}$$

$$l''(\beta) = -\frac{n-c}{\beta^2} < 0$$

pa funkcija l poprima maksimum u $\beta = \frac{n-c}{\sum_{i=1}^{n} t_i}$. Dakle, MLE za parametar β je broj necenzuriranih primjera kroz ukupno vrijeme promatranja.

Linearna regresija

Kod podatkovne analize linearna regresija zajedno s metodom procjene najmanjih kvadrata jedan je on najkorištenijeg pristupa. Pristup linearne regresije ne može se direktno primijeniti za analizu preživljenja budući da nedostaju vremena događaja od interesa za cenzurirane primjere, no postoje metode koje su razvijene za rješavanje tog problema.
Tobit regresija jedna je od najranijih pokušaja proširenja linearne regresije s Gausovom razdiobom za analizu cenzuriranih primjera. Neka je y^* linearno ovisna o parametru β i vrijedi $y^* = X\beta + \epsilon$, pri čemu je $\epsilon \sim N(0, \sigma^2)$, normalno distribuirana greška.

$$y_i = \begin{cases} y_i^*, & \text{ako vrijedi } y_i^* > 0. \\ 0, & \text{inače.} \end{cases} \tag{2.6}$$

Obzirom na y^* parametri modela mogu biti aproksimirani koristeći metodu maksimalne vjerodostojnosti pri čemu je kompleksnost $O(NP^2)$.
Buckley-James regresija aproksimira vrijeme doživljenja pomoću Kaplan-Meier metode i onda na model linearne regresije (AFT) primjeni aproksimirana vremena s obzirom na vremena doživljenja necenzuriranih opažanja.
Metode sažimanja Linearna regresija uz korištenje metode sažimanja s različitim regularizatorima za višedimenzionalne cenzurirane podatke učinkokovita je metoda za obradu cenzuriranih podataka davajući im različite težine. Primjeri metoda sažimanja su Ridge (hrbatna), Lasso *(engl. least absolute shrinkage and selection)* i elastična mreža *(eng. elastic net)*.

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Ridge i lasso metode sažimaju koeficijente regresije stavljanjem kazne na njihovu veličinu. Parametar $\lambda \geq 0$ je parametar složenosti koji upravlja iznosom sažimanja, odnosno veća vrijenost λ predstavlja veći iznos sažimanja. Razlika između ove dvije metode je u izražavanju kazne, ridge regresija ima kaznu izraženu u 2 - normi (euklidska norma) dok je kod lasso metode u 1-normi.
Weighted regresija može biti korisna kod analize doživljenja kako bi se naglasak stavio

na informacije koje više doprinose modelu. Weighted regresija minimizira težinsku sumu kvadrata $\sum_{i=1}^{n} \omega_i(y_i - X_i\beta)^2$. Kompleksnost te regresije je $O(NP)$.

Model ubrzanog vremena otkazivanja

Kod parametarskih, cenzuriranih regresijskih modela pretpostavljamo da vrijeme doživljenja svih obzervacija prati neku specifičnu distribuciju s linarnom vezom između varijabli i vremena doživljenja ili logaritma vremena doživljenja. Posebno, ako je veza između logaritma vremena doživljenja T i varijabli linearna tada koristimo parametarski model, **model ubrzanog vremena otkazivanja (AFT model)**.

Definicija 2.3.6. Neka je X=$(X_1, ..., X_n)$ opaženi uzorak za slučajnu varijablu X, $\beta = (\beta_1, \beta_2, ..., \beta_n)$ koeficijenti , model ubrzanog vremena dan je izrazom

$$\log(T) = \beta^T X + \sigma\epsilon \text{ ili } T = e^{\beta^T X} e^{\sigma\epsilon} \tag{2.7}$$

pri čemu je σ parametar skaliranja, a ϵ označava slučajnu pogrešku opažanja. Pripadna funkcija doživljenja je:

$$S_x(t) = S_0(e^{-\beta^T X} t) \tag{2.8}$$

pri čemu je $S_0(t)$ osnovna funkcija doživljenja *(eng. baseline)* .

U većini slučajeva pretpostavlja se da slučajna pogrška ϵ prati jednu od distribucija iz tablice 2.6 . U tom slučaju preživljenje ovisi i o kovarijatama i o pripadajućoj distribuciji. Faktor ubrzanja, koji je po pretpostavci konstantan kroz vrijeme T, dan je formulom $e^{-\beta X}$ i on na vrijeme doživljenja djeluje usporavajuće, odnosno ubrzavajuće. Ukoliko je $e^{-\beta X} < 1$, vrijeme doživljenja prolazi sporije, za $e^{-\beta X} > 1$ prolazi brže, a za $e^{-\beta X} = 1$ ide normalnom brzinom.

Eksponencijalni i Weibull modeli mogu biti parametrizirani kao modeli proporcionalnog hazarda ili model ubrzanog vremena otkazivanja. Kod drugih parametarskih modela (log-normalan, gamma) parametri nisu nužno osnovni hazard baseline multipliciran konstantom te mogu biti izraženi samo kao model ubrzanog vremena otkazivanja.

Poglavlje 3

Metode strojnog učenja

Jednu od opisnih definicija za strojno učenje dao je Tom M. Mitchell: „Kažemo da program uči na temelju iskustva E obzirom na neki skup zadatka T i evaluacijskom mjerom P, ako se njegov učinak na zadacima T poboljšava sa iskustvom E evaluirano mjerom P." Disciplina strojnog učenja je bazirana na pitanju kako konstruirati kompjuterski program koji automatski poboljšava iskustvo. Iako se strojno učenje počelo teoretski razvijati već 1950-ih godina, posljednjih nekoliko godina razvijeno je puno uspješnih primjena, detekcija prevare kod kreditnih transakcija, informacijskog filtriranja preferencija za pojedinog korisnika, autonomna vozila koji sama voze na autocestama. Razvili su se algoritmi te je napredovala teorija.
Tri su glavne podjele strojnog učenja:

(1) **nadzirano učenje** (*eng.supervised learning*): Podaci su dani u parovima (ulaz, izlaz)=(x, y) te treba pronaći procjenitelj $\hat{y}$ = f(x). Cilj je napraviti model koji će raditi predikcije na još neviđenim (novim) primjerima (klasifikacija, regredija, predikcija)

(2) **nenadzirano učenje** (*eng.unsupervised learning*): dani su podaci bez ciljne vrijednosti, a treba naći pravilnost u podacima (grupiranje, otkrivanje outlier-a, smanjenje dimenzionalnosti)

(3) **učenje s podrškom** (*eng. reinforcement learning*): učenje optimalne strategije na temelju pokušaja s odgođenom nagradom (obično vezano uz učenje sekvenci akcija - igre)

U analizi preživljenja glavni izazov za primjenu metoda strojnog učenja su cenzurirani podataci te predviđanje događaja od interesa u odnosu na vrijeme.

3.1 Stablo preživljenja

Stablo preživljenja je metoda klasifikacijskih i regresijskih stabala (CART), prilagođena za cenzurirane podatke. Osnovna intuicija primjenjena na model stabala je rekurzivna podjela podataka s obzirom na odeđeni kriterij tako da su slični podaci u odnosu na događaj od interesa smješteni na isti čvor staba. Kao i kod stabla odlučivanja, čvorovi predstavljaju varijable, grane vrijednosti varijable, a listovi stabla odluke, označimo ih s F. Primarna razlika stabla preživljenja u odnosu na osnovno stablo odlučivanja je odabir kriterija podjele. Smatra se da je čvor „čist" ako svi subjekti u čvoru prežive identičan inteval vremena. Koristi se log-rank statistika kao kriterij koja procjenjuje razliku preživljenja.

Bagging stabla preživljenja

Bagging stablo preživljenja zapravo je kombinacija bagginga (*bootstrap aggregating*) i stabla preživljenja. Prvo se uzme B bootstrap poduzorka skupa za učenje (to su poduzorci dobiveni uzrokovanjem s ponavljanjem). Nakon toga se za svaki podskup radi stablo preživljenja bazirano na svim značajkama. Čvorovi se dijele rekurzivno korištenjem klasifikatora koji maksimizira razliku preživljenja između susjednih čvorova. Na novu primjer se primijeni bootstrap agregirana funkcija doživljenja.

3.2 Ansambli

Kombiniranje predikcija više modela koji su napravljeni s istim ili različitim algoritmom na istim ili različitim podacima s ciljem poboljšavanja predikcije u odnosu na jedan model naziva se ansambli.

Varijacijom početnih točaka često je moguće konstruirati ansambl koji daje bolju aproksimaciju, pogotovo za male količine podataka. Kako bi se izbjegla nestabilnost korištenja jedne metode, Leo Breinman predložio je kreiranje modela ansamblom korištenjem metoda bagging i slučajna šuma.

Slučajna šuma preživljenja

Slučajne šume su korištene primarno za klasifikacijske i regresijske probleme Proširenje metode slučajne šume na desno cenzurirane podatke je od značajne važnosti. Podaci vezani za analizu preživljenja oslanjaju se na ograničavajuće pretpostavke poput proporcionalnosti hazarda. Kod korištenja parametarskih modela neliearan odnos varijabli mora biti modeliran s različitim transformacijama. Korištenjem ad hoc pristupa poput stepwise regresije provjerava se nelinearnost. Osim toga, problematično je i identificiranje inte-

rakcija između varijabli. Ti problemi automatski su riješeni korištenjem metode slučajne šume. Slijedi opis algoritma slučajnih šuma prikazanog na slici 3.1:

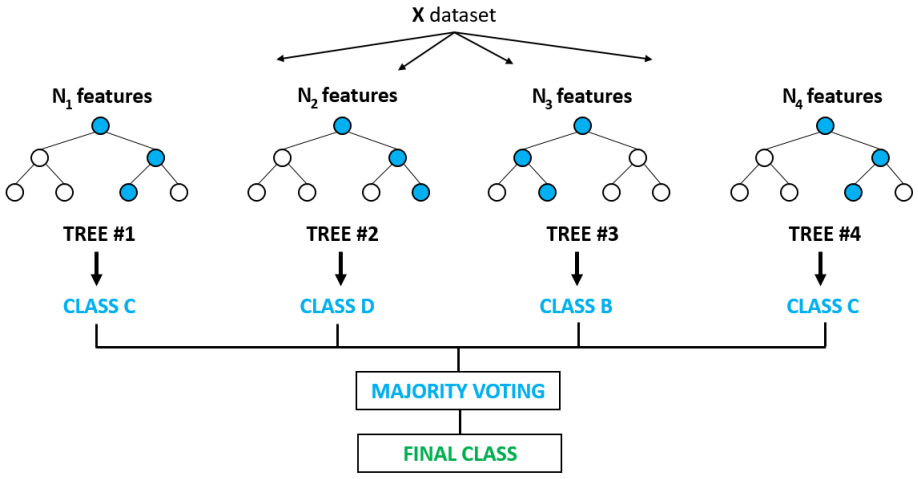
1. Prvi korak je razdvojiti skup podataka na B boostrap podskupova. Svaki podskup ne koristi u prosjeku 37% podataka, kojeg nazivamo „podaci izvan skupa" (*eng. out of bag*), skraćeno OOB.

2. Za svaki bootstrap podskup trenira se posebno stablo. Na svakom čvoru se slučajnim obabirom izabire p varijabli koji su kandidati za čvorove. Čvor se dijeli korištenjem varijable koja maksimizira razliku preživljenja između susjednih čvorova.

3. Stablo se generira do trenutka kada čvor ima manje od d_0 događaja od interesa.

4. Za svako stablo računa se vrijednost kumulativne funkcije hazarda te se uzima prosjek tih vrijednosti svih stabala.

5. Koristeći OOB podatke računa se greška predviđanja na ansamblu kumulativnih funkcija rizika.



Slika 3.1: Slučajna šuma, izvor:[10]

Na kraju stablo preživljenja završava u listovima F za koje vrijedi kriterij da sadrže minimum d_0 događaja od interesa. Neka su $(T_{1,h}, \delta_{1,h}), ..., (T_{1,h}, \delta_{n(h),h})$ parovi vremena preživljenja i oznake je li podatak cenzuriran za list h $\in F$. Uzorak i je desno cenzuriran u vremenu $T_{i,h}$ ako je $\delta_{i,h} = 0$, a $\delta_{i,h} = 1$ označava događaj od interesa u vremenu $T_{i,h}$.

Neka su $t_{1,h} < t_{2,h} < ... < t_{N(h),h}$ N(h) različitih vremena, $m_{l,h}$ je broj događaja od interesa u vremenu $t_{l,h}$ i $r_{l,h}$ broj osoba u riziku u vremnu $t_{l,h}$. Kumulativna funkcija hazarda za list h procijenjena je Nelson-Aalen procjeniteljem

$$\hat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{m_{l,h}}{r_{l,h}}.$$

Bootstrap i OOB ansambl kumulativne funckije hazarda

Opisana kumulativna funkcija hazarda izračunata je za svako pojedino stablo. Kako bi se izračunala kumulativna funkcija hazarda za ansambl uzima se prosjek svih B stabala. Opisani su OOB i bootstrap procjenitelji. Svako stablo u šumi je napravljeno iz nezavisnog bootstrap podskupa. Neka je $I_{i,b} = 1$ ako je i OOB od stabla b, a inače je $I_{i,b} = 0$. Neka je $H_b^*(t|x)$ kumulativna funkcija hazarda za stablo napravljeno od b bootstrap podskupa. OOB procjenitelj za kumulativnu funkciju od i je

$$H_e^{**}(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b^*(t|x_i)}{\sum_{b=1}^{B} I_{i,b}}.$$

Odnosno, to je prosjek nad bootstrap poduzorkom gdje je i OOB. Ekvivalentno, $H_e^{**}(t|x_i)$ se može izračunati tako da se OOB poduzorak spušta po stablu preživljenja iz podataka unutar skupa (*eng. in-bag*) bootstrap podataka pamteći i-ti list i pripadnu kumulativnu funkciju preživljenja. Na kraju se uzima prosjek svih vrijednosti.
Bootsrap ansambl kumulativne funkcije hazarda za i dan je s

$$H_e^*(t|x_i) = \frac{1}{B} \sum_{b=1}^{B} H_b^*(t|x_i).$$

Uočimo da se u formuli koriste sva stabla preživljenja, a ne samo ona gdje je i OOB.

Boosting

Boosting algoritmi su široko korištena metoda ansambla za kombiniranje slabih modela dajući težinu pojedinim modelima prema njihovoj točnosti, takvi algoritmi smanjuju pristranost i varijancu konačnog modela. Iterativno prilagođava odgovarajuće definirane reziduale bazirano na algoritmu gradijentnog boostinga. Algoritam je prilagođen da minimizira težinsku funkciju rizika $\hat{\beta}_{\tilde{U},X} = \text{argmin}_\beta \sum_{i=1}^{N} \omega_i(\tilde{U}_i - h(X_i|\beta))$, pri čemu je $\tilde{U}_i$ rezidual $\tilde{U}_i = -\frac{\partial L(y,\phi)}{\partial \phi}\Big|_{\phi=f_m(\hat{X}_i)}$, β vektor parametara, a $h(\cdot|\hat{\beta_{U,X}})$ procjenitelj.

3.3 Metoda potpornih vektora

Metoda potpornih vektora (*eng. support vector machines, SVM*) je uspješan model nadziranog učenja primjenjiv primarno na klasifikacijske probleme, ali može biti i prilagođen za regresijske. Metoda potpornih vektora nalazi hiper-ravninu koja ima najveću marginu razdvajanja klase (udaljenost između „kritičnih" točaka, najbliže blizu plohi razvajanja). Ova metoda je uspješno prilagođena i za analizu preživljenja. Za taj pristup koriste se kernel funkcije s kojima transformacijama svaku točku mapiramo u neki novi prostor. Na taj način omogućeno je pretvaranje neseparabilnih problema u separabilne. Neka je dan skup podataka za treniranje (x_i, y_i) gdje je $x_i \in R^n$, a y_i klasa primjera (-1 ili 1).

Metoda potpornih vektora traži rješenje optimizirajućeg problema:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

pri čemu vrijede uvjeti:

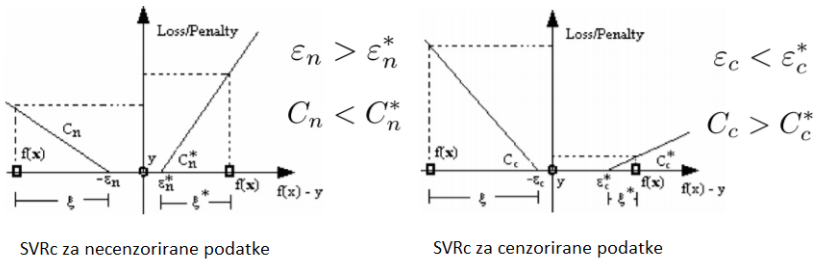
$$|y_i - (w\phi(x_i) + b)| \leq \varepsilon + \xi$$

$$\xi_i, \xi_i^* \geq 0, i = 1, ..., n$$

Pritom je w normalizirani vektor okomit na hiperravninu, C regularizacijski parametar, ξ su tzv. *slack variables* koje mjere pogrešku u točki (x_i, y_i), a predstavljaju pomak hiperravnine duž vektora w.

$$|\xi|_\varepsilon = \begin{cases} 0 & ;|\xi_i| \leq \varepsilon \\ |\xi_i| - \varepsilon & ;\text{inače} \end{cases}$$

Zbog cenzuriranih podataka ne možemo izravno primijeniti regresiju metode potpornih vektora, nego moramo modificirati funkciju gubitka/troška (*loss function*).



Slika 3.2: Prikaz metode potpornih vektora, izvor: [6]

Aktivno učenje

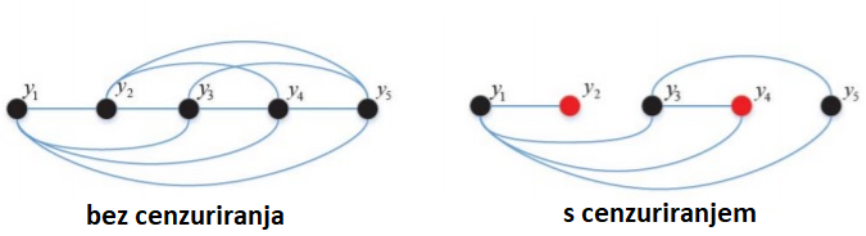
Aktivno učenje primjenjeno na cenzurirane podatke može biti jako korisno za analizu preživljenja jer mišljenje stručnjaka može biti korišteno u modelu. Mehanizam aktivnog učenja dopušta da model prvo izabere podskup uzorka učeći na limitiranom označenom setu, a onda ispita stručnjaka za doprinos u obliku oznake preživljenja prije uzimanja training seta.

3.4 Mjere uspješnosti modela kod analize preživljenja

Zbog prisutnosti cenzuriranih podataka za analizu preživljenja nisu primjenjive neke standardne mjere uspješnosti modela poput srednje kvadratne pogreške i mjere R^2.

C indeks

Uobičajeni način mjerenja uspješnosti kod analize preživljenja je proučavajući relativan rizik događaja za različite istance u odnosu na apsolutno vrijeme preživljenja. To se radi mjerenjem C indeksa, odnosno mjere suglasnosti (*eng. concordance index*) . Vrijeme preživljenja dvije istance može biti uspoređivano za dva slučaja, prvi, kada su oba vremena cenzurirana i drugi kada je promatrano vrijeme necenzurirane instance manje nego cenzurirano vrijeme. Na slici 3.3 je prikazano grupiranje za slučaj bez cenzuriranja za koji postoji 10 mogućih parova, te s cenzuriranim vremenima za koji postoje 6 mogućih parova. Usporedba (uparivanje) cenzuriranih istanci može biti samo s ranijim necezuriranim istancama (na slici par y_2 i y_1).



Slika 3.3: Primjer grupiranja za izračun C indeksa, ($y_1 < y_2 < y_3 < y_4 < y_5$) ,izvor:[10]

Neka je dan par (i,j) s vremenima preživljenja t_i i t_j i T_i, T_j predviđenim vremenima :

- par (i,j) je suglasan ako vrijedi $t_i > t_j$ te $T_i > T_j$
- par (i,j) nije suglasan ako vrijedi $t_i > t_j$ te $T_i < T_j$.

Vjerojatnost suglasnosti c=$P(\hat{T}_i \geq \hat{T}_j|T_{ij})$ mjeri suglasnost između odnosa stvarne vrijednosti i predviđene. U praksi postoje više načina računanja C indeksa:
(1) kada je rezultat modela omjer hazarda (na primjer Cox model):

$$\hat{c} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I[X_I\hat{\beta} > X_j\hat{\beta}]$$

, pri čemu $i, j \in$ {1,...,N}, *num* broj usporedivih parova te $I[\cdot]$ indikatorska funkcija, a $\hat{\beta}$ procjenjeni parametri Cox modela.
(2) za modele koji procjenjuju vrijeme preživljenja C indeks je oblika:

$$\hat{c} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I[S(\hat{y}_j|X_j) > S(\hat{y}_i|X_i)]$$

, pri čemu su S(·) procjenjene vjerojatnosti preživljenja. Slično kao i kod AUC, vrijednost C indeksa 1 predstavlja najbolji model predviđanja, a vrijednost 0.5 slučajno predviđanje. Kako bi se mogla testirati uspješnost modela tijekom različitog vremena promatranja definiran je C indeks za određeni period $(0, t*)$ kao težinski prosjel AUC vrijednosti za sva moguća promatrana vremena. Vremenski zavisan AUC za specifično vrijeme preživljenja t može se izračunati pomoću formule:

$$AUC(t) = P(\hat{y}_i < \hat{y}_j|y_i < t, y_j > t) = \frac{1}{num(t)} \sum_{i:y_i<t} \sum_{j:y_j>t} I(\hat{y}_i < \hat{y}_j)$$

, gjde je $t \in T_s$,tj. unutar skupa svih vremena preživljenja, a $num(t)$ jednak je broju usporedivih parova u periodu $(0, t^*)$. C indeks za period $(0, t*)$ dan je formulom:

$$c_{t*} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I(\hat{y}_i < \hat{y}_j) = \sum_{t \in T_s} AUC(t) \cdot \frac{num(t)}{num}$$

Srednje kvadratna pogreška

Srednje kvadratna pogreška (MAE *eng. mean apsolute error*) za analizu preživljenja definirana je kao prosjek razlika između predviđenih i stvarnih vremena događaja od interesa. Neka je s y_i (I=1,...,N) označeno stvarno promatrano vrijeme, a s $\hat{y}_i$ procjenjeno vrijeme, vrijedi:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} (\delta_i|y_i - \hat{y}_i|).$$

Za cenzurirane podatke vrijedi $\delta_i = 0$ pa takvi podaci ne ulaze u sumu. Procjena uspješnosti MAE može biti korištena samo kod modela koji predviđaju vrijeme preživljenja poput AFT modela.

Brier mjera

Mjera brier (*eng. brier score*, skraćeno BS) nazvana po Glenn W. Brier, mjera je razvijena za potrebe predviđanja pogreški prilikom modeliranja vremenske porognoze. Može biti korištena samo za modele čiji je rezultat vjerojatnost, tj. izlazna varijabla modela mora biti unutar intervala [0,1], a suma svih mogućih izlaznih varijabli 1. Kada proučavamo binarne modele dužine uzorka N, pri čemu je za svaki X_i (i=1,2,...,N), predviđena vjerojatnost u vremenu t označena s $\hat{y_i}(t)$,a stvarna vjerojatnost $y_i(t)$ (0 ili 1) tada je fromula za računanje BS dana s:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i(t) - y_i(t)]^2$$

Brier score je proširen za potrebe mjerenja uspješnosti modela za analizu preživljenja kada je izlazna varijabla ili binarna ili kategorijska. Za cenzurirane podatke u uzorku doprinos Breir score-u je težinski u odnosu na distribuciju. Neka je

$$\omega_i(t) = \begin{cases} \delta_i/G(y_i) & \text{ako je } y_i \leq t \\ 1/G(y_i) & \text{ako je } y_i > t \end{cases}$$

odnosno ω_i je težina za $i-tu$ instancu i procjenjena je koristeći Kaplan-Meier procjenitelj za funkciju doživljenja cenzuriranih podataka G za dani set podataka $(X_i, y_i, 1-\delta_i)$, $i = 1, ..., N$. BS mjera je prilagođena na sljedeći način:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\omega_i[\hat{y}_i(t) - y_i(t)]^2$$

Pretpostavlja se da su vremena preživljenja necenzuriranih i cenzuriranih podataka nezavisna. BS može biti proširen za računanje na intervalu

$$IBS = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} BS(s)ds$$

U praksi se gornji integral aproksimira numeričkom integracijom.

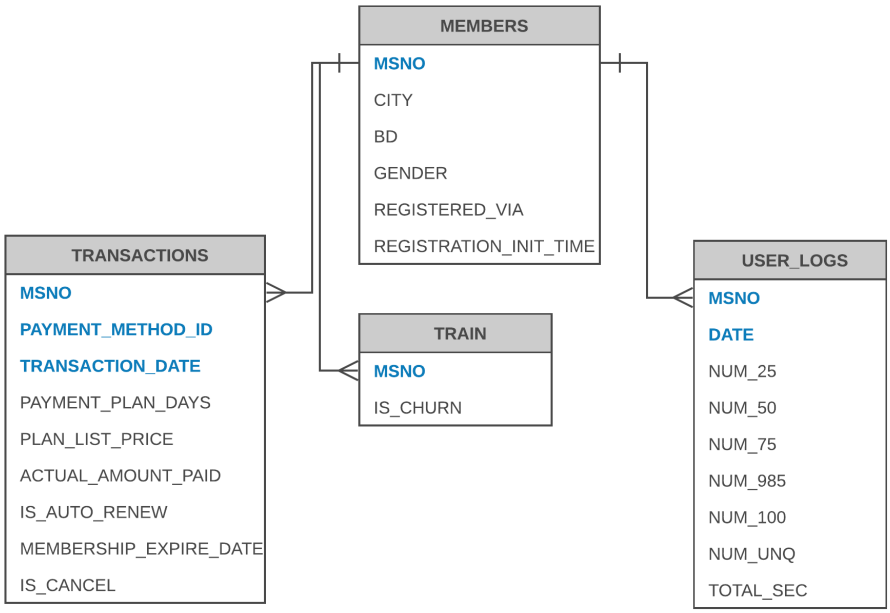
Poglavlje 4

Analiza preživljenja za Kkbox podatke

Nakon što je opisan teorijski dio analize preživljenja, napisano ću primijeniti. Podaci na kojima je napravljena analiza preuzeti su s Kaggle natjecanja. Kaggle je platforma podatkovnih znanstvenika i entuzijasta koja je primarno zajednica gdje se razmjenjuju iskustva, kodovi i znanje. Kaggle je 2017. godine premašio 1.000.000 korisnika iz 194 zemalja i najveća je i najrazličitija platforma takvog tipa. Kaggle natjecanja sadrže probleme iz raznolikih područja znanosti i ekonomije na kojima se primjenjuje podatkovna znanost. Natjecanja privlače veliki broj ljudi zbog izazovnih zadataka, ali i zbog novčanih nagrada za najbolje timove. Nerijetko najbolje rangirani timovi nakon završetka natjecanja podijele sa zajednicom dijelove koda i način pristupa problemu što omogućava lako učenje novima u tom području i usavršavanje iskusnijih.
Naziv natjecanja je „WSDM - KKBox's Churn Prediction Challenge - *Can you predict when subscribers will churn?*", a cilj je predvidjeti koji korisnici će odustati od pretplate. Objavljeno je 2018. godine s nagradnim fondom od 5.000 $ i privuklo je ukupno 575 timova.
Za razliku od inicijalno zadanog problema, primjenit ću analizu doživljenja i naglasak staviti na pitanje: **kada korisnici odustaju od pretplate**. Kkbox je glazbena aplikacija popularna u Aziji, koja između ostalog, sadži više od 30 milijuna pjesama žanra Asia-Pop. Kroz različite pakete pretplate milijunima ljudi nude različite opcije pri čemu im je za što bolje poslovanje jako bitno predvidjeti odlazak *eng. churn* korisnika. Trenutno KKBOX koristi različite metode od kojih je jedna i analiza preživljenja kako bi predvidjeli ukupno vrijeme pretplate za svakog korisnika. Cilj im je saznati zašto korisnici odlaze te kako ih mogu zadržati (npr. potencijalnim posebnim ponudama za određene korisnike, popustima ...).

Opis seta podataka

Za analizu su dostupna 4 seta podataka koji su učitani u 4 dokumenta u csv fromatu. U svakom dokumentu nalaze se različiti tipovi podataka o korisnicima.
Members je set podataka o pojedinom korisniku kao što su grad (city), dob (bd), spol (gender), način registracije (registered via) i početno vrijeme registracije (registration init time). Btino je napomenuti da su u dokumentima sadržani podaci o puno više korisnika u odnosu na one kojima je naznačen churn te one za koje se očekuje predikcija.
U dokumentu **transactions** nalaze se transakcije svih korisnika od 1.1.2015. do 31.3.2017. Transakcija je zapravo pretplata opisana tipom plaćanja (payment method id), planom otplate (payment plan days, plan list price) , plaćenim iznosom (actual amount paid) te podatkom o isteku članstva (memebership expiry date). Nadalje, u tablici se nalazi podatak koji označava je li pretplata automatski obnovljena (is auto renew = 1) ili ne, is cancel označava je li korisnik sam otkazao pretpatu (1) ili nije (0). No, otkazivanje pretplate ne znači nužno churn korisnika. Najveći broj transakcija jednog korisnika je 244, a ukupno dokument ima 22.978.755 zapisa.



Slika 4.1: Podaci o korisnicima u obliku tablica

Svakodnevni logovi korisnika, odnosno dnevna aktivnost zabilježena je u tablici user logs. U njoj se nalaze podaci kao oznaka koliko pjesama je korisnik u tom danu poslušao

do 25% (num 25) , koliko jedinstvenih pjesama je poslušao (num unq) te ukupno vrijeme aktivnosti u sekundama (total secs). Budući da su aktivnosti dnevne taj dokument je najveći i veličina dokumenta je preko 30GB, što predstavlja veliko opterećenje za memoriju računala. U dokumentu **train** nalaze se korisnici za koje je označena churn vrijednost pomoću kkbox algoritma za ožujak 2017. godine. Kasnije ću pokazati da je algoritam nepravilno primjenjen te sama kreirati podatak o churnu pa zadnji dokument neću koristiti.

4.1 Programski alati

Prvo su opisani programski alati korišteni za analizu. Nakon toga opisujem na koji način sam obradila sirove podatke i izradila konačan set podataka.

Oracle SQL Developer, OracleXE

Oracle SQL Developer je inregrirano developersko okruženje (*eng. Integrated development environment (IDE)* za korištenje SQL i PL/SQL jezika u Oracle bazi podataka. Unutar programa nalazi se dio za pisanje i pokretanje upita i skripti, DBA konzola za upravljanje bazom, kompletno rješenje za modeliranje, te migracijska platforma za povezivanje s nekom drugom bazom. Besplatna verzija OracleXE (reducirana verzija baze Oracle 11g) dostupna je za javno korištenje uz neka ograničenja: prostor za korisničke podatke je ograničen na 11 GB, koristi najviše 1GB RAMa, dostupna je samo jedna instanca po poslužitelju te koristi samo jedan CPU u višeprocesorskom okruženju. Nakon instalacije na tu bazu sam se spojila s programom Oracle SQL Developer. SQL je jezik za rad s relacijskom bazom podataka, sadrži 3 tipa naredbi: DDL za kreiranje objekata, DML za manipuliranje nad podacima te DQL za postavljanje upita. Navedene programe koristila sam za učitavanje csv dokumenata te kako bih dobila početni uvid u podatke. Nakon toga podatke sam spajala po primarnom ključu (msno) te radila agregacije i grupacije.

Python

Python je programski jezik najviše korišten u području podatkovne znanosti, prva verzija napravljena je 1991. godine i sam jezik je ime dobio po seriji: ”Monty Python’s Flying Circus”. Razlozi odabira Pythona kao programske podrške najviše leže u tome što je uvelike prilagođen analizi preživljenja i većina metoda je implementirana u jednoj od širokog spektra biblioteka. Biblioteke u Pythonu uvelike doprinose funkcionalnosti samog jezika, a neke od biblioteka korištenih za izradu ovog diplomskog rada su Dask, Pandas, Numpy, Plotly te Lifelines. Lifelines je biblioteka predodređena za analizu preživljenja i nju sam koristila za izradu modela, Dask sam koristila za učitavanje velikih csv datoteka (datoteke

logova koju zbog memorije nisam mogla učitati izravno u Oracle SQL Developer), a Plotly za izradu interaktivnih vizualizacija.

Churn - odlazak klijenta

Definicija odlaska klijenta je zamršena zbog specifičnog modela KKBOX pretplate. Budući da većina korisničkih pretplata traje 30 dana, puno korisnika obnavlja pretplatu svaki mjesec. Ključ za određivanje odlaska klijenta su 3 varijable: datum transakcije (transaction date), datum isteka članstva (membership expiration date) i poništenje pretplate (is cancel). Poništenje pretplate se dogodi kada osoba (korisnik) sam odluči odustati od tog oblika pretplate no to ne znači da je osoba prestala biti korisnik. To može napraviti zbog različitih razloga, na primjer da pređe na drugi način pretplate. Kriterij za oznaku odlaska klijenta je da nema aktivnih pretplata 30 dana nakon otkazivanja ili 30 dana nakon što mu je prijašnja pretplata istekla.
U dokumentu transakcije nalaze se zapisi po korisniku koji odgovaraju obnovi pretplate ili otkazivanju. Prvo je bilo potrebno preraditi dokument tako da se uz pojedinu pretplatu nalazi datum kada je sklopljena, datum isteka te oznaka je li na kraju klijent tu pretplatu otkazao te ako je otkazao promijeniti datum isteka. Kada je dobiveni transformirani dokument korištenjem SQL upita usporedbom datuma dobivena je oznaka is churn za svaki pojedini zapis i period. Za jednog korisnika dobiven je podatak koliko puta je ukupno prestajao biti korisnik aplikacije, a kako bi se ustanovio trend dodane su i varijable poput broja churn-a predzadnje i zadnje godine te mjeseca.

	MSNO	PAYMENT_METHOD_ID	PAYMENT_PLAN_DAYS	PLAN_LIST_PRICE	ACTUAL_AMOUNT_PAID	IS_AUTO_RENEW	TRANSACTION_D...	MEMBERSHIP_EXPIRE_DATE	IS_CANCEL
1	UgTB51AJZjdcYkWYaQF0/8...	35	7	0	0	0	08-JUN-15	15-JUN-15	0
2	UgTB51AJZjdcYkWYaQF0/8...	29	30	180	180	1	16-FEB-17	16-MAR-17	0
3	UgTB51AJZjdcYkWYaQF0/8...	29	30	180	0	1	16-MAR-17	18-MAR-17	0
4	UgTB51AJZjdcYkWYaQF0/8...	29	30	180	0	1	18-MAR-17	20-MAR-17	0
5	UgTB51AJZjdcYkWYaQF0/8...	29	30	180	0	1	20-MAR-17	22-MAR-17	0
6	UgTB51AJZjdcYkWYaQF0/8...	29	30	180	0	1	22-MAR-17	24-MAR-17	0
7	UgTB51AJZjdcYkWYaQF0/8...	29	30	180	180	1	24-MAR-17	24-MAR-17	1

Slika 4.2: Primjer transakcija jednog pretplatnika

Na slici 4.2 možemo vidjeti podatke o transkacijama za jednog pretplatnika. Pretplatniku je prvi put članstvo aktivirano 6.6.2015. godine i imao je tjedan dana besplatno korištenje aplikacije. Nakon toga ukupno je 5 puta obnovio pretplatu i jednom je (zadnjom transakcijom) otkazao pretplatu. Članstvo mu ističe 24.3.2017. pa se zbog toga ne smatra churn korisnikom (nije prošlo 30 dana neaktivnosti).

Na slici 4.3 prikazane su transakcije korisnika koji je dva puta prestajao biti korisnik. Može se vidjeti da je 11.6.2015. prekinuo pretplatu i opet postao pretplatnikom 18.11.2016. Nakon toga nije imao niti jednu transakciju te mu je članstvo isteklo 25.10.2016. zbog toga za tog korisnika je označen događaj od interesa, tj. churn.

	MSNO	PAYMENT_METHOD_ID	PAYMENT_PLAN_DAYS	PLAN_LIST_PRICE	ACTUAL_AMOUNT_PAID	IS_AUTO_RENEW	TRANSACTION_D...	MEMBERSHIP_EXPIRE_DATE	IS_CANCEL
1	CJP9mEjfHqvX8dBW...	37	31	149	149	1	10-JAN-15	10-FEB-15	0
2	CJP9mEjfHqvX8dBW...	37	31	149	149	1	10-FEB-15	10-MAR-15	0
3	CJP9mEjfHqvX8dBW...	37	31	149	149	1	10-MAR-15	10-APR-15	0
4	CJP9mEjfHqvX8dBW...	37	31	149	149	1	11-APR-15	10-MAY-15	0
5	CJP9mEjfHqvX8dBW...	37	0	0	149	1	11-MAY-15	10-JUN-15	0
6	CJP9mEjfHqvX8dBW...	37	30	149	149	1	11-JUN-15	10-JUN-15	1
7	CJP9mEjfHqvX8dBW...	35	7	0	0	0	18-NOV-16	25-NOV-16	0

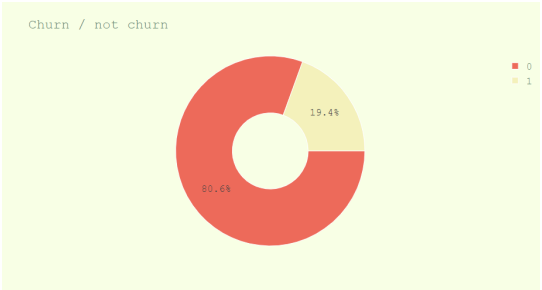
Slika 4.3: Primjer transakcija churn pretplatnika

4.2 Opisna statistika

Prvi korak pri analizi podataka će biti opisna statistika kako bi se dobio uvid u podatke i vezu između njih. Ukupno je za potrebe modeliranja izdvojeno N=1.019.356 zapisa, odnosno korisnika. Na početku u setu podataka bilo je 57 varijabli od kojih je jedino varijabla koja označava spol korisnika imala nepoznate (NA) vrijednosti i to njih 557.133. Zbog tako velikog broja podataka koji nedostaju varijabla spol je odmah izbačena iz daljnje analize. Podaci po značenju mogu biti svrstani u tri dijela:

- podaci o transakcijama (agregirano za zadnja 3 mjeseca aktivnosti, za cijeli period te vezani za zadnju transakciju)
- podaci o korisniku (spol, dob, grad, tip registracije)
- agregirani podaci o logovima korisnika (agregirani podaci o broju poslušanih pjesama do nekog postotka, ukupan broj poslušanih sekundi)

U setu podataka nalaze se 52 numeričke i 4 kategorijske varijable. Kategorijske varijable su: BD (dob), REGISTERED VIA (6 različitih načina registracije), PAYMENT METHOD ID LAST (33 različita načina plaćanja) te CITY (21 grad). Podaci su prikupljeni do 3. mjeseca 2017. godine što znači da su desno cenzurirani, varijabla TRAJANJE označava ukupno vrijeme aktivnosti korisnika izraženo u danima, a varijabla koja označava događaj od interesa (vrijednost 1) ili cenzurirane podatke (vrijednost 0) je IS CHURN.



Slika 4.4: Omjer cenzuriranih i necenzuriranih podataka

Iz slike 4.4 možemo primjetiti da set podataka nije balansiran. Samo 19,4% korisnika (njih 198.163) ima označen događaj od interesa (churn) na kraju razdoblja. Nebalansirani set podataka može prestavljati problem kod predikcije te se može dogoditi da modeli slabije prepoznaju churn.

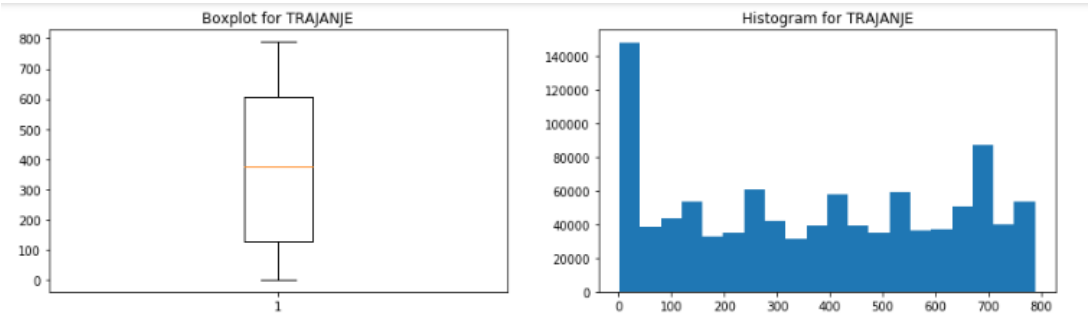
Sažetak varijabli

feature	count	mean	std	min	25%	50%	75%	max
TRAJANJE	1019356	372.9	252.1	1	127	374	608	789
UKUPAN_BR_TRANS	1019356	11.7	8.4	1	3	11	19	65
UKUPAN_BR_OTKAZIVANJA	1019356	0.2	0.6	0	0	0	0	11
UKUPAN_BR_CHURNA	1019356	0.7	0.9	0	0	0	1	9
UKUPAN_BR_AUTO_RENEW	1019356	9.5	9	0	0	8	18	61
BR_POPUSTA	1019356	0.7	1.4	0	0	0	1	46
UKUPAN_BR_TRANS_P2Y	1019356	2.8	4.2	0	0	0	5	47
UKUPAN_BR_OTKAZIVANJA_P2Y	1019356	0.1	0.2	0	0	0	0	7
UKUPAN_BR_CHURNA_P2Y	1019356	0.2	0.4	0	0	0	0	5
UKUPAN_BR_TRANS_PY	1019356	6.1	4.8	0	1	7	12	46
UKUPAN_BR_OTKAZIVANJA_PY	1019356	0.1	0.3	0	0	0	0	7
UKUPAN_BR_CHURNA_PY	1019356	0.3	0.5	0	0	0	0	5
CHRUN_P2M	1019356	0	0.1	0	0	0	0	1
COUNT_P2M	1019356	0.7	0.5	0	0	1	1	5
CANCEL_P2M	1019356	0	0.1	0	0	0	0	2
CHRUN_PM	1019356	0	0.2	0	0	0	0	1
COUNT_PM	1019356	0.7	0.5	0	0	1	1	6
CANCEL_PM	1019356	0	0.2	0	0	0	0	2
CHRUN_LM	1019356	0.2	0.4	0	0	0	0	3
COUNT_LM	1019356	1	0.2	0	1	1	1	6
CANCEL_LM	1019356	0.1	0.2	0	0	0	0	3
IS_CHURN	1019356	0.2	0.4	0	0	0	0	1
NUM_LOGS	1019356	210.8	213.7	0	28	139	342	821
NUM_25_AVG	1019356	6	7.3	0	2.3	4.2	7.4	1409.8
TOTAL_SECS_AVG	1019356	404000704	113256071168	0	2713.5	4758.2	7562.7	4990084423
NUM_50_AVG	1019356	1.5	2	0	0.6	1.1	1.9	341
NUM_75_AVG	1019356	0.9	1	0	0.4	0.7	1.1	240.2
NUM_985_AVG	1019356	0.9	1.5	0	0.4	0.7	1.1	159.2
NUM_100_AVG	1019356	22.1	22.5	0	9.1	17.2	28.2	1584
NUM_UNQ_AVG	1019356	23.1	18	0	11.9	19.8	29.9	1308.2
NUM_25	1019356	4.9	7.9	0	0.8	2.8	6.1	1710.9
NUM_50	1019356	1.2	2.2	0	0.1	0.7	1.5	452
TOTAL_SECS	1019356	5197.6	6235.7	0	1282	3767.3	6786.7	477913.1
NUM_75	1019356	0.7	1.1	0	0	0.5	1	208.5
NUM_985	1019356	0.8	1.7	0	0	0.5	1	189.5
NUM_100	1019356	19.6	25.6	0	3.9	13.6	25.4	1922.5
NUM_UNQ	1019356	19.9	20.6	0	5.7	15.7	27.3	1560.2
P_NUM_25	1019356	3.9	6.9	0	0	1.9	5.1	1409.8
P_NUM_50	1019356	1	1.9	0	0	0.5	1.3	243
P_TOTAL_SECS	1019356	4379.7	6071	0	0	2945.6	6067	410530.5
P_NUM_75	1019356	0.6	0.9	0	0	0.3	0.9	95
P_NUM_985	1019356	0.6	1.4	0	0	0.3	0.9	253.6
P_NUM_100	1019356	16.6	24.7	0	0	10.4	22.6	2224.7
P_NUM_UNQ	1019356	16.7	20.2	0	0	12.5	24.7	1308.2
BD	1019356	13.1	20.1	-5978	0	0	26	2016
PAYMENT_METHOD_ID_COUNT	1019356	1.3	0.6	1	1	1	1	9
PAYMENT_PLAN_DAYS_S	1019356	360.3	240.8	1	150	390	540	2821
PLAN_LIST_PRICE_S	1019356	1577.4	1137.4	0	594	1485	2533	12857
ACTUAL_AMOUNT_PAID_S	1019356	1637.9	1213.2	0	594	1485	2688	12857
PLAN_LIST_PRICE_LAST	1019356	162.8	218.3	0	99	149	149	2000
ACTUAL_AMOUNT_PAID_LAST	1019356	162.3	218.5	0	99	149	149	2000

Slika 4.5: Tablica svih varijabli i pripadna osnovna statistika

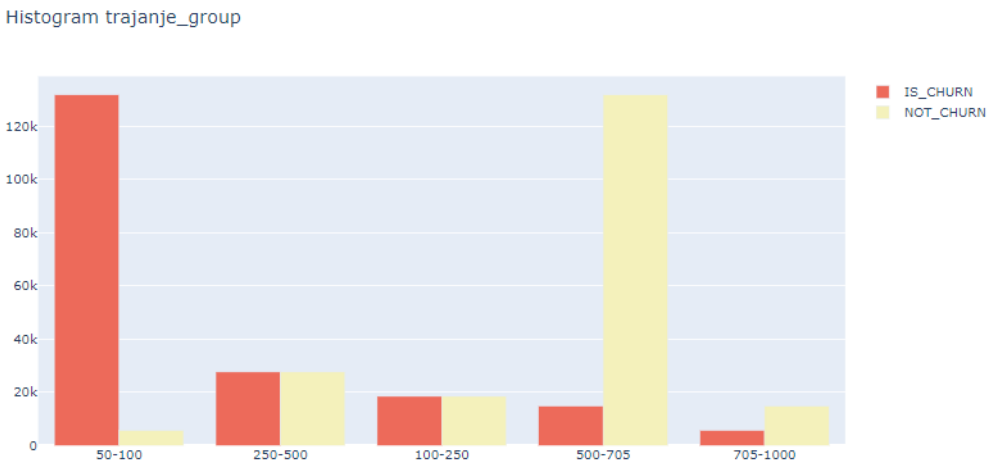
Iz osnovnih statistika provedenih nad varijablama prikazanih u tablici 4.5 mogu se primijetiti određeni outlieri, tj. vrijednosti koje odstupaju od ostalih. Outlieri su prisutni kod prosjeka poslušanih sekundi (TOTAL SECS AVG). Varijablu BD izbacujem iz daljnje analize jer ima ukupno 563.396 krivo unesenih vrijednosti (vrijednosti koje su manje od 13,

a veće od 65). Također, u tablici vidimo maksimalne i minimalne vrijednosti, maksimalno trajanje pretplate je 789 dana, a najveći broj pretplata 65.



Slika 4.6: Histogram i box-plot varijable trajanje

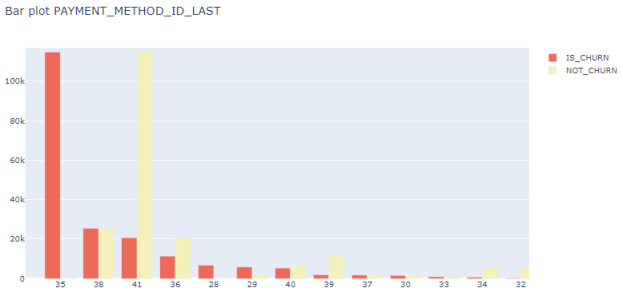
Na slici 4.6 prikazana je distribucija trajanja članstva, najviše je novih korisnika, koji su članovi do 100 dana. Međutim, takvi korisnici imaju i najveći broj churna, čak 131.704 (66% churn korisnika). Skoro pa jednako toliko korisnika je cenzurirano s trajanjem između 500 do 705 dana.



Slika 4.7: Trajanje u odnosu na churn

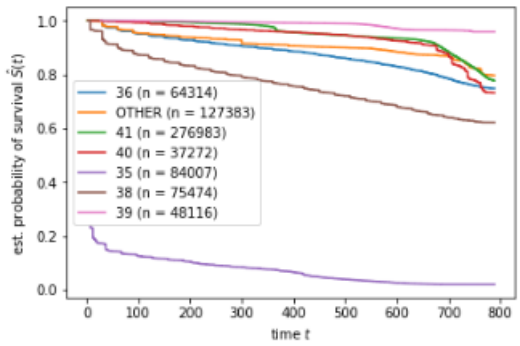
Kako vrijeme raste vidimo da ima sve manje churn korisnika. Veliki broj korisnika ima članstvo dulje od 500 dana i tada je zabilježeno najmanje odlazaka.

Osim varijable vremena zanimljivo je pogledati koja je ovisnost događaja od interesa s raznim oblicima transakcija. Na slici 4.8 je vidljivo da je metoda s kojom je pretplatnik platio zadnju transakciju prošlog mjeseca povezana sa churnom za metodu plaćanja 41, najkorišteniju metodu gdje je postotak churna jako malen i za metodu plaćanja 35 koju su koristili gotovo svi churn pretplatnici.



Slika 4.8: Metoda plaćanja u odnosu na churn

Kaplan Meier funkciju preživljenja u ovisnosti o metodi plaćanja možemo vidjeti na slici 4.9. Može se uočiti da korisnici koju su zadnju transakciju platili s metodom 38 imaju 80% šansu preživljenja kada im je vrijeme pretplate 30 dana.

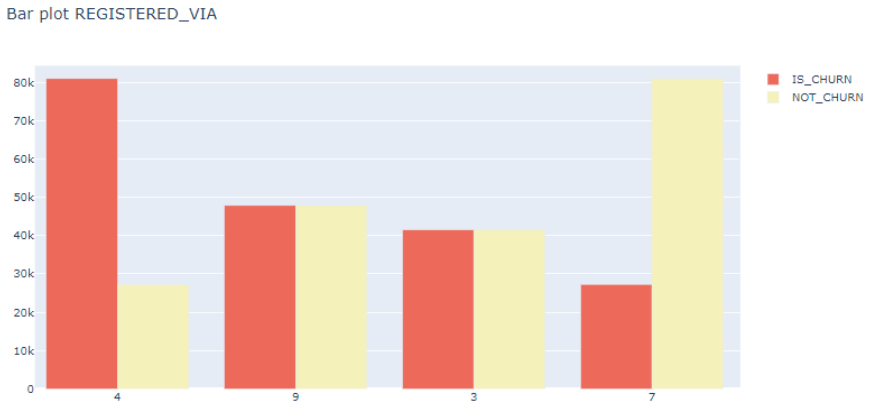


Slika 4.9: Metoda plaćanja u odnosu na churn

Najveći postotak korisnika „preživljava" kada im je metoda plaćanja 39. Kada se bolje analiziraju podaci vidi se da pretplatnici koji imaju oznaku plaćanja 35 imaju i planiran iznos plaćanja 0. Vjerojatno se radi o nekoj vrsti jednokratne besplatne pretplate. Ako je to način da se privuku dugotrajni korisnici ili zadrže stalni vidi se da dugoročno nije uspješno. Za modeliranje je možda korisno uzeti interakciju varijabli metode plaćanja i

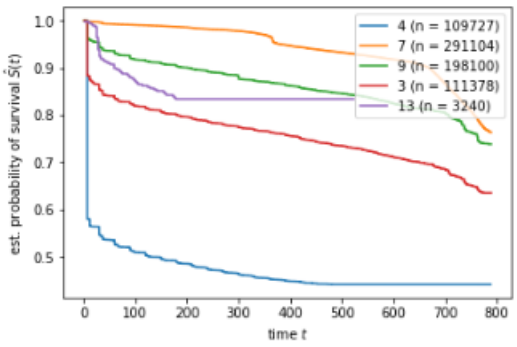
planiranog iznosa. Budući da nisu dane informacije o kojim metodama se točno radi, tj. značenje šifre, ne može se zaključiti ništa više od ovoga.

Dok se podaci o transakcijama mijenjaju tijekom vremena, za podatak o metodi registracije važno je naglasiti da je to podatak jedinstven za svakog korisnika i ne mijenja se u odnosu na vrijeme. Za korisnike koji su registrirani metodama 4 i 7 postoji najveća razlika između cenzuriranih vrijednosti i churna.



Slika 4.10: Metoda registracije u odnosu na churn

Na slici 4.11 je vidljivo da čak 50% korisnika koji su se registrirali metodom 4 ima označen churn nakon 100 dana. Funkcije preživljenja za metode 7, 9 i 3 povećanjem vremena padaju skoro paralelno, a najveća razlika churna je u prvih 200 dana.



Slika 4.11: KM prikaz za metodu registracije

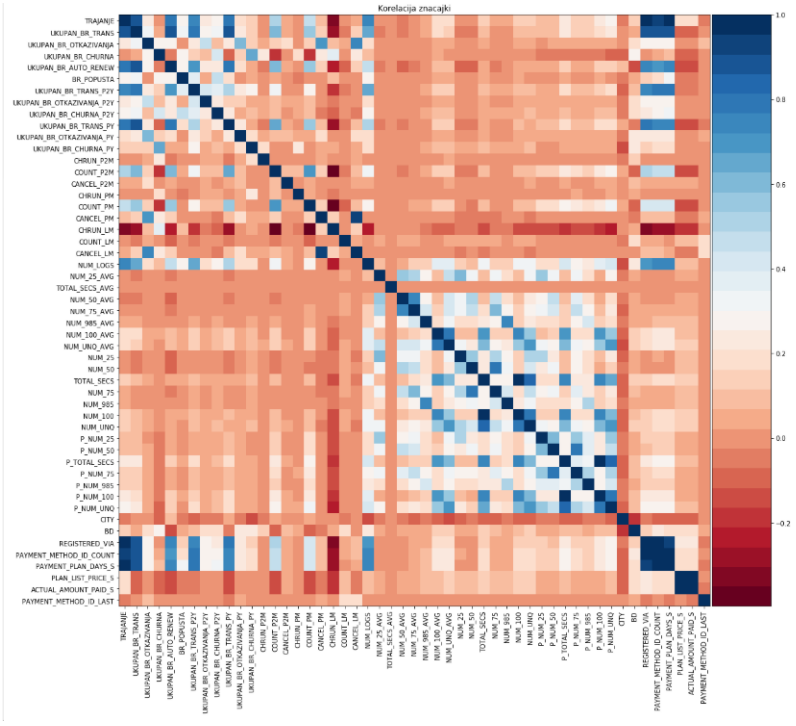
Deskriptivna analiza je jako korisna kako bi se vidjela povezanost varijabli te dobila ideja za modeliranje i važnost varijabli. Međutim, deskriptivna analiza nije dovoljna kako

bi se donijeli neki značajni zaključci koji bi nam poslužili za predikciju niti usporedile sve varijable.

Korelacija varijabli

Kako bi se izmjerio linearan stupanj povezanosti varijabli korišten je Pearsonov koeficijent korelacije. Koreliranost ne utječe na pouzdanost modela (bar ne na setu na kojem treniramo model) nego na pojedine prediktore. Procijenjeni parametri i p vrijednosti mogu se značajno promijeniti ako napravimo male promjene u modelu ili podacima.

Matrica korelacije prikazuje koreliranost između svih varijabli. Ukoliko je vrijednost Pearsonovog koeficijenta korelacije veća od nule varijable su pozitivno korelirane, ako je manja od nule negativno, a ako je jednak nuli onda varijable nisu korelirane.



Slika 4.12: Matrica korelacije

Na slici 4.12 je matrica korelacije pri čemu su prikazane vrijednosti Personovog koeficjenta u bojama na skali od -0.6 do 1. Postoji jako puno koreliranih varijabli poput plan list price last i actual amount paid last te sam za takve varijable radila razne transformacije ili ih izbacila iz modela. Kod iznosa actual i planed sam napravila novu varijablu razlike i odbacila planed, na taj način očuvana je informacija, a smanjena korelacija.

Nakon što su izbačene jako korelirane varijable i stvorene nove, prije stvaranja modela potrebno je odabrati „najznačajnije" varijable. Odabir podskupa varijabli je potreban jer smanjuje vrijeme kreiranje modela (računalne restrikcije) i poboljšava interpretabilnost modela te pridonosi boljoj predikciji.

Odabir varijabli za modeliranje

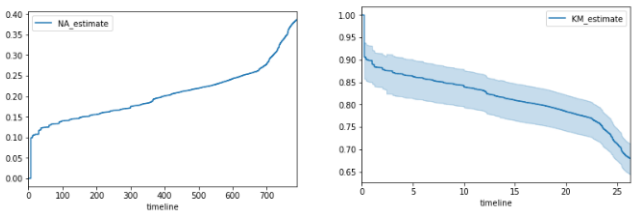
Za odabir varijabli korištene su dvije metode koje su neovisne o modelu, χ^2-statistika te Pearsonov koeficijent (ovaj put u odnosu na ciljnu varijablu) i 4 metode koje su ovisne o modelu. Korištene metode koje su ovisne o modelu su Lasso, slučajne šume, XG Boost te rekurzivna eliminacija. Inicijalno sam zadala broj varijabli za svaku metodu $k = 30$. **Rekurzivna eliminacija** bira varijable rekurzivno, svaki put smanjujući skup varijabli. Prvo se evaluira važnost varijabli na cijelom setu varijabli te se iteracijama izbacuje varijabla s najlošijom važnosti. Za rekurzivnu eliminaciju korištena je logistička regresija. **Lasso metoda** koristi L1 normu težina kao regularizator. Iduće su korištene metode slučajne šume koja evaluira važnost varijable u odnosu na točnost stabla kada ona je/nije u modelu/stablu te na kraju i XG Boost.

Na temelju rezultata izabrala sam 19 najznačajnijih varijabli te njih koristila za daljnje modeliranje.

4.3 Modeliranje

Neparametarske metode

Neparametarske metode Kaplan-Meier i Nelson Aalen prikazuju funkcije preživljenja i kumulativne funkcije hazarda. Mogu se koristiti kako bi se vidio odnos različitih vrijednosti značajke u odnosu na preživljenje, kao što smo iskoristili za gornje usporedbe 4.9. Generalno, preživljenje korisnika KKBOX aplikacije vidljivo je na slici 4.13, pri čemu je *KM* procjena napravljena na mjesečnoj, a *NA* na dnevnoj razni.



Slika 4.13: Procjena funkcije kumulativnog hazarda i funkcije doživljenja

Cox model

Cox model, poznat kao i model proporcionalnih hazarda jedan je od najkorištenijih modela za analizu preživljenja. Jako bitna pretpostavka kod navedenog modela je proporcionalnost hazarda, odnosno da su funkcije hazarda jednake do na množenje konstantom. To implicira da je funkcija hazarda za bilo koja dva subjekta u bilo kojem trenutku proporcionalna. Ako se prekrši ova pretpostaka ne može se koristiti osnovni Cox model i potrebne su modifikacije. Na podacima iz KKBOX seta nije bilo moguće primijeniti osnovni Cox model. Naime pretpostavka nije bila narušena samo za tri varijable s rezultatom Cox modela 0.72 (C indeks).
Na slici 4.14 prikazan je Cox model s odličnim C-indeksom, čak **0.96**, i značajnim varijablama. Međutim, Cox pretpostavka o proporcionalnosti hazarda nije zadovoljena pa model nije relevantan.

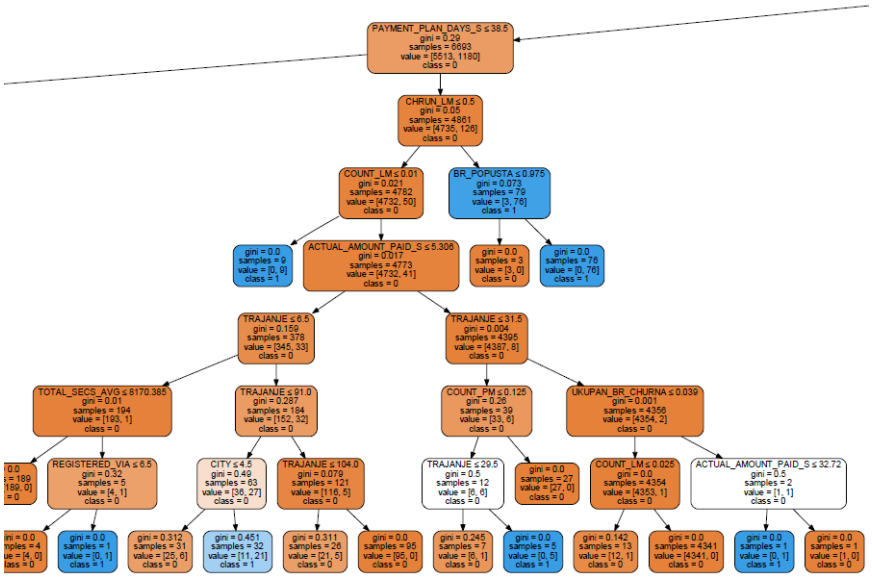
Stratification je postupak koji se radi kada varijable ne zadovoljavaju pretpostavku o

```
Iteration 6: norm_delta = 0.00000, step_size = 1.0000, ll = -1567288.73867, newton_decrement = 0.00000, seconds_since_start =
1.4Convergence completed after 6 iterations.
<lifelines.CoxPHFitter: fitted with 713549 observations, 574879 censored>
      duration col = 'TRAJANJE'
         event col = 'IS_CHURN'
number of subjects = 713549
  number of events = 138670
partial log-likelihood = -1567288.74
  time fit was run = 2019-09-02 11:51:49 UTC

---
                        coef exp(coef)  se(coef)  coef lower 95%  coef upper 95%  exp(coef) lower 95%  exp(coef) upper 95%
P_NUM_UNQ              -0.04      0.96      0.00           -0.04           -0.04                 0.96                 0.96
ACTUAL_AMOUNT_PAID_LAST -0.01      0.99      0.00           -0.01           -0.01                 0.99                 0.99
UKUPAN_BR_AUTO_RENEW    -0.17      0.84      0.00           -0.17           -0.17                 0.84                 0.84
UKUPAN_BR_OTKAZIVANJA    0.81      2.26      0.00            0.81            0.82                 2.25                 2.27
COUNT_P2M               -0.88      0.42      0.01           -0.90           -0.86                 0.41                 0.42

                              z      p  -log2(p)
P_NUM_UNQ              -144.25 <0.005       inf
ACTUAL_AMOUNT_PAID_LAST -278.51 <0.005       inf
UKUPAN_BR_AUTO_RENEW    -267.52 <0.005       inf
UKUPAN_BR_OTKAZIVANJA    296.02 <0.005       inf
COUNT_P2M               -105.40 <0.005       inf
---
Concordance = 0.96
Log-likelihood ratio test = 502283.72 on 5 df, -log2(p)=inf
```

Slika 4.14: Cox model

proporcionalnosti hazarda. Takve varijable ostaju u modelu, ali se ne procjenjuje njihov doprinos. Set podataka se dijeli u l manjih setova s obzirom na jedinstvene vrijednosti stratifikacijskih varijabli. Svaki od njih ima posebnu osnovnu funkciju hazarda, ali zajedničke koeficijente. Za stratifikacijske varijable se ne procjenjuju koeficijenti u Cox modelu. Za Cox model u kojem nisu narušene pretpostavke o proporcionalnosti hazarda korištene su dvije stratifikacijske značajke tip registracije i način zadnjeg plaćanja. C-index tog Cox modela je **0.85**. Budući da se model oslanja samo na pet značajki upitna je kvaliteta predikcije.

Stablo odlučivanja



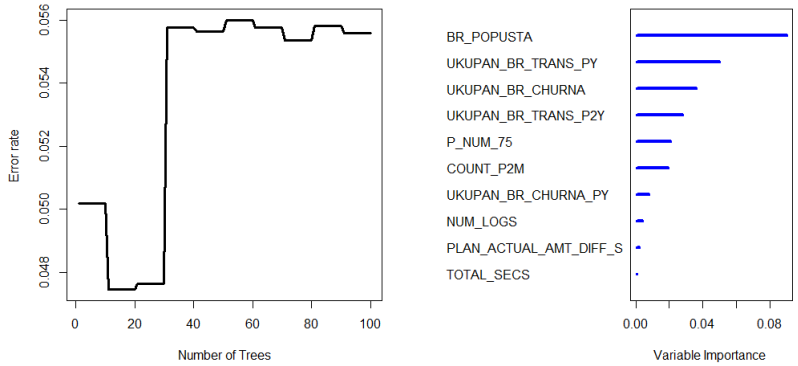
Slika 4.15: Jedna grana osnovnog stabla odlučivanja

Za analizu podataka može se koristiti i stablo odlučivanja posebno implementirano za analizu preživljenja. Za ovu metodu korištena je Python biblioteka *sksurv*. Kod implementacije stabla potrebno je u argumente staviti posebno set podataka za treniranje te posebno pripadno trajanje i varijablu is churn. Metoda je bazirana na regresijskim stablama s Cox gradijentom. C-indeks te metode je **0.79**.

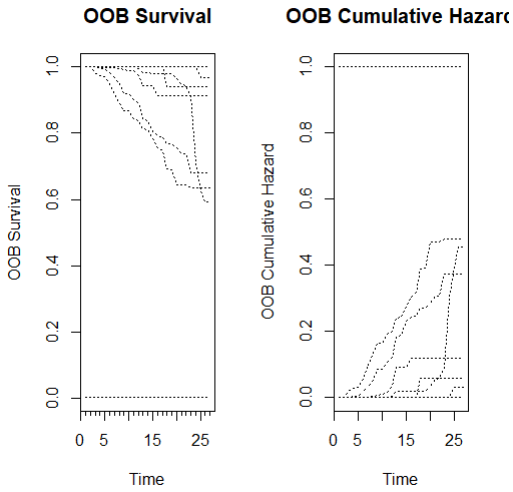
Slučajna šuma preživljenja

Algoritam slučajne šume preživljenja implementiran je unutar nekoliko biblioteka u Pythonu, no paket *randomForestSRC* u programskom jeziku R je daleko najpopularniji i najkorišteniji. Zbog memorijske zahtjevnosti algoritma u Pythonu za modeliranje su korištena oba jezika, R za pronalaženje najboljih parametara za gradnju šume te Python za implementaciju [1]. Zbog karakteristika algoritma i jačine kompjutera na kojima je izvođen, korišteni su podskupovi podataka s najviše 500.000 zapisa i 19 varijabli. Stabla odlučivanja inače mogu podnjeti veći skup podataka i veći broj varijabli jer model stabla sam „probire" najznačajnije varijable za izgradnju modela. Kod konfiguracije modela slučajne šume najbitniji su parametri poput dubine stabla, *learning ratea* te minimalanog broja u konačnom listu. Ostali parametri su broj varijabli slučajno izabranih za kandidate čvorova, metoda

bootstrapa itd. Set za treniranje dobiven je uzimanjem reprezentativnog seta podatka temeljenim na jednakoj distribuciji seta u odnosu na trajanje i broj cenzuriranih podataka. Najveći C-indeks (mjera uspješnosti) dobiven za algoritam slučajne šume bio je **0.87**. Dobiven je na skupu podataka koji je sadržavao svih 19 značajki te s brojem stabala 100, uzorku veličine 400.000 i maksimalne veličine lista 15. Na slikama su prikazana svojstva šume na manjim setovima podataka.



Slika 4.16: Varijable po relativnoj važnosti u modelu te ovisnost greške modela i broja stabala



Slika 4.17: Predviđena funkcija preživljenja i kumulativnog hazarda za 10 subjekata iz OOB uzorka

4.4 Zaključak i otvorena pitanja

Rezultati prikazanih metoda na KKBOX setu po pokazateljima uspješnosti su dobra početna točka za daljnje modeliranje. Koristeći samo mjeru C-indeks kao mjeru uspješnosti modela iz dobivenih rezultata se može zaključiti da je najuspješniji model slučajnih šuma. Zbog restrikcija, vezanih za jačinu te memoriju kompjutera nije izvršena pretraga najboljih parametara modela (za slučajnu šumu broj stabala, čvorova ...) niti su isprobavane neuronske mreže za analizu preživljenja. Također, svi prikazani modeli su mogu dalje razvijati kako bi se dobila i veća uspješnost. Podaci dani za KKBOX natjecanje nisu u potpunosti iskorišteni te je moguće razviti još značajnih prediktora, posebno istražujući razne interakcije između varijabli. Iako osnovna pretpostavka o proporcionalnosti hazarda nije ispunjena moguće je napraviti razne modifikacije Cox modela za vremenski zavisne varijable. Samim razvijanjem područja podatkovne znanosti, zasigurno će u budućnosti biti puno više naglaska na razvijanju jakih modela za predikciju na podacima preživljenja.

Bibliografija

[1] *Algoritam slučajna šuma preživljenja u Pythonu*, https://github.com/Wrymm/Random-Survival-Forests.

[2] E. H. Blackstone M. S. Lauer H. Ishwaran, U. B. Kogalur, *Random survival forests*, (2008), https://arxiv.org/pdf/0811.1645.pdf.

[3] A. Jazbec, *Odabrane statističke metode u biomedicini, PMF-MO, nastavni materijali 2016.*

[4] Kaggle, *WSDM - KKBox's Churn Prediction Challenge*, (2017), https://www.kaggle.com/c/kkbox-churn-prediction-challenge.

[5] P. G. Karadeniz i I. Ercan, *Examining Tests For Comparing Survival Curves With Right Censored Data*, Statistics in Transition New Series **18** (2017), br. 2, 311–328, https://ideas.repec.org/a/exl/29stat/v18y2017i2p311-328.html.

[6] F. M. Khan i V. B. Zubek, *Support vector regression for censored data (SVRc): a novel tool for survival analysis*, (2008).

[7] David G. Kleinbaum, *Survival Analysis : A Self -Learning Text*, Springer, 2012, ISBN 0-387-94543-1.

[8] Tom M. Mitchell, *Machine learning*, 1., McGraw-Hill, 1997, ISBN 0-07-115467-1.

[9] Kishore J. M.K. Goel, Khanna P, *Understanding survival analysis: Kaplan-Meier estimate*, (2010), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/.

[10] Y. LI Chandan K. P. Wang, V. Tech, *Machine Learning for Survival Analysis: A Survey*, (2017.), https://arxiv.org/abs/1708.04649.

[11] M. P. Barman R. Saikia, *A Review on Accelerated Failure Time Models*, (2017), https://www.ripublication.com/ijss17/ijssv12n2_15.pdf.

[12] Nikola Sarapa, *Teorija vjerojatnosti*, Školska knjiga, 2002, ISBN 978-953-0-30816-9.

[13] Steve Selvin, *Statistical Analysis of Epidemiologic Data*, 2004, ISBN 0-19-517280-9, third edition.

[14] T. Šmuc, *Strojno učenje, PMF-MO, nastavni materijali 2018.*

Sažetak

Analiza preživljenja iznimno je korisna kada želimo predvidjeti vrijeme događaja od interesa. Postoji sve više podataka prikladnih za analizu preživljenja i sve veća potreba za analiziranje istih. Postoji cijeli skup statističkih metoda, a sve se više razvijaju i metode strojnog učenje prilagođene za glavno svojstvo podataka o preživljenju, cenzuriranje. Dan je pregled statističkih metoda analize preživljenja i metoda strojnog učenja. Metode su uspoređene te je za dan pregled prednosti i mana. Navedne su i opisane mjere uspješnosti modela kod analize preživljenja. U ovome radu su opisane metode primijenjene na primjeru iz stvarnog svijeta, na podacima glazbene aplikacije.

Summary

Survival analysis is extremely useful when we want to predict the time of the event of interest. There are a lot of developed statistical methods and machine learning is getting more and more adapted to the main problem with survival data, censorship. In this thesis is given a review of conventional survival analysis methods and various machine learning methods for survival analysis. Methods were compared and the main advantages and disadvantages of each were specified. Using specialized evaluation metrics for prediction performance was described. In this thesis main methods for analyzing such data are applied to the music application data.

Životopis

Rođena sam 21. prosinca 1993. godine u Zagrebu, gdje sam završila Osnovnu školu August Šenoa i Devetu gimnaziju. Po završetku srednje škole, 2012. godine upisujem preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu, koji završavam 2016. godine stekavši naziv sveučilišna prvostupnica matematike. Iste godine upisujem diplomski studij Matematičke statistike koji završavam ovim radom. Tijekom studiranja bila sam članica dvije studentske udruge eSTUDENT i BEST koje su upotpunile iskustvo studiranja. Sudjelovala sam na dva studentska natjecanja vezana za rješavanje stvarnih poslovnih slučajeva (*Case Study Competition*) i pobijedila na oba.