Analiza

"Women's Tennis Association Matches"

WTA matches from 2000-2016

skupa podataka

Maja Gavrilović 489/2017

Septembar 2019

Sažetak:

U ovom radu predstavljamo rezultate dobijene analizom teniskih mečeva na turnirima u periodu od 2000 do 2016 godine. Prvo ćemo se upoznati sa opisom skupa podataka, dok kasnije stavljamo akcenat i bavimo se predikcijom mečeva koja omogućava korisniku da za dva tenisera dobije ishod meča. Jedan od izazova je bio postupak pretprocesiranja o kome ćemo detaljnije u nastavku...

Sadržaj:

Opis i vizualizacija skupa podataka

“Women’s tennis association matches” je skup podataka koji sadrži rezultate teniskih mečeva u periodu 2000 do 2016 godine. U istraživanje je uključeno i sakupljeno 2876 instanci. Ovaj skup koristićemo za razvoj prediktivnog modeliranja teniskih mečeva i za statističko istraživanje. Skup se sastoji od sledećih fajlova:

• wta_matches_2000.csv: schema koja uključuje podatke iz 2000 godine.

• wta_matches_2001.csv: schema koja uključuje podatke iz 2001 godine.

• wta_matches_2002.csv: schema koja uključuje podatke iz 2002 godine.

• wta_matches_2003.csv: schema koja uključuje podatke iz 2003 godine.

• wta_matches_2004.csv: schema koja uključuje podatke iz 2004 godine.

• wta_matches_2005.csv: schema koja uključuje podatke iz 2005 godine.

• wta_matches_2006.csv: schema koja uključuje podatke iz 2006 godine.

• wta_matches_2007.csv: schema koja uključuje podatke iz 2007 godine.

• wta_matches_2008.csv: schema koja uključuje podatke iz 2008 godine.

• wta_matches_2009.csv: schema koja uključuje podatke iz 2009 godine.

• wta_matches_2010.csv: schema koja uključuje podatke iz 2010 godine.

• wta_matches_2011.csv: schema koja uključuje podatke iz 2011 godine.

• wta_matches_2012.csv: schema koja uključuje podatke iz 2012 godine.

• wta_matches_2013.csv: schema koja uključuje podatke iz 2013 godine.

• wta_matches_2014.csv: schema koja uključuje podatke iz 2014 godine.

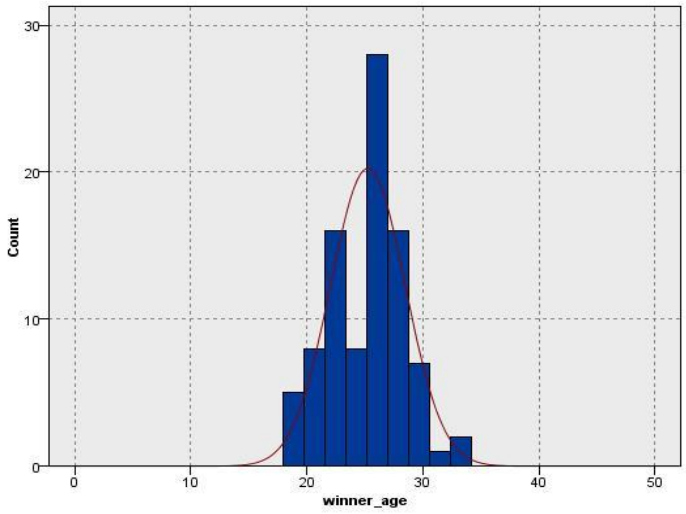
• wta_matches_2015.csv: schema koja uključuje podatke iz 2015 godine.

• wta_matches_2016.csv: schema koja uključuje podatke iz 2016 godine.

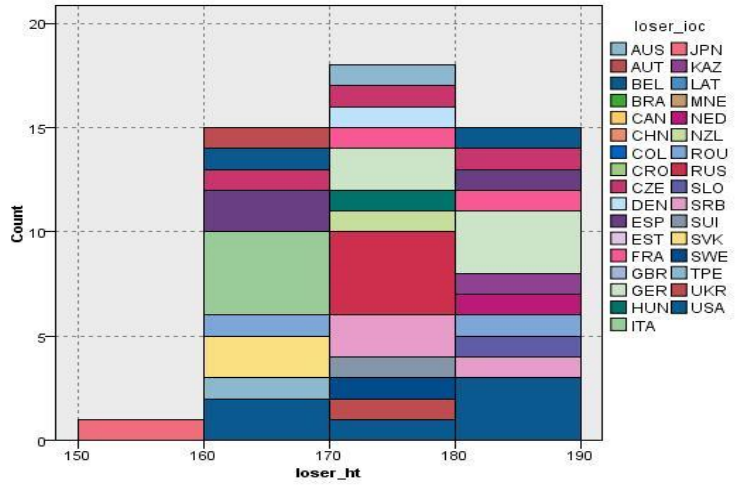
Neki od glavnih korišćenih atributa i njihove vizualizacije će biti prikazane u nastavku:

Shema 2016:

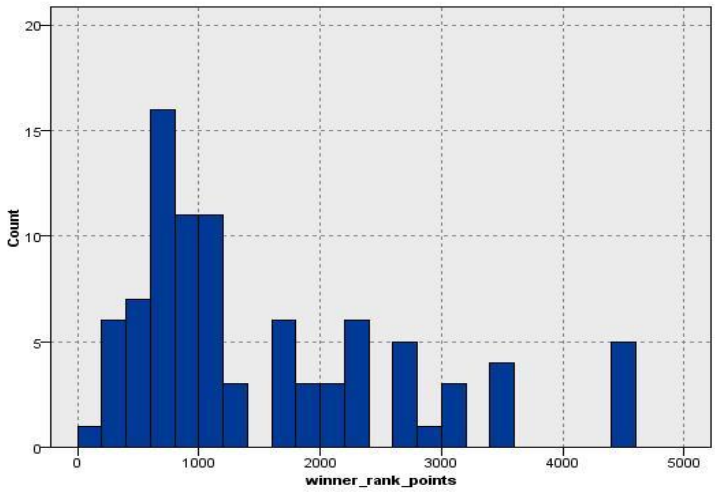
• winner_age – starost teniserki koje su odnele pobedu, u odnosu na normalnu raspodelu



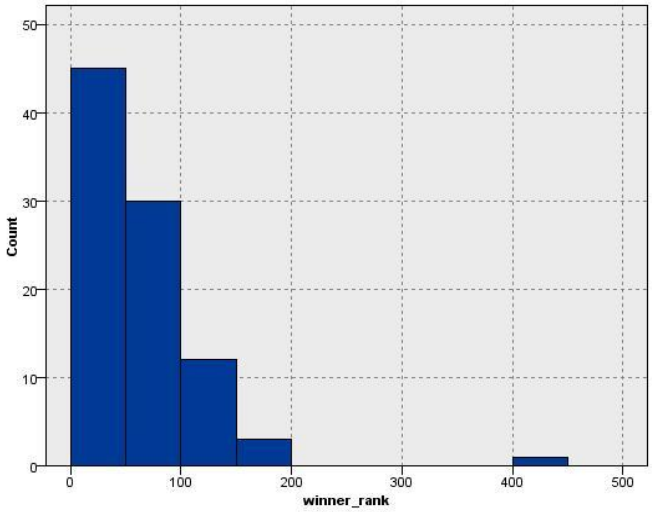
• looser_ht-visina teniserki koje su dozivele poraz prikazane po zemlji porekla



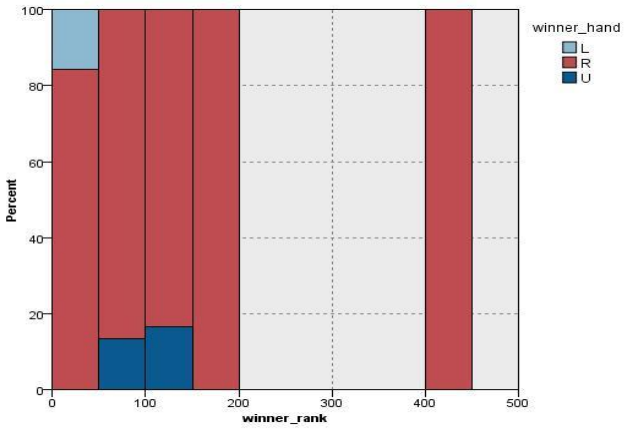
• winner_rank_points-rangiranje na osnovu broja poena u osvojenim mečevima, na skali do 5000



- winner_rank-rangiranje teniserki koje su odnele pobedu



- Isti histogram prikazan u procentima kojom rukom teniserke igraju



Korišćeni alati

Za obradu tj. klasifikovanje i vizualizaciju korišćen je IBM SPSS, dok je za pretprocesiranje korišćen jezik Python i biblioteke numpy I pandas.

Pretprocesiranje

Jedan od težih delova za pretprocesiranje bilo je nalaženje statistika. Iz rada kog navodim[1] sam izvukla statistike za računanje. Sve sto sam dobila iz kolona je dato kroz apsolutne brojeve. Nek od kolona koje sam izračunala su procenat osvojenih poena na prvom servisu itd.. Ovde nailazimo do problema. U ovom data setu mi nemamo output kolonu, zapravo nemamo target.. Vec znamo unapred podatke o pobedniku I gubitniku meča... Morala sam da se dosetim 2 stvari. Ideja jedan je bila da napravim nove kolone po referenci na rad koji sam malopre navela.

1 https://www.researchgate.net/publication/330883807_Predicting_the_Outcome_of_a_Tennis_Tournament_Based_on_Both_Data_and_Judgments?fbclid=IwAR1ImVd6uXKdbYvQMoEfQ2y60bSYea-BNf-fHi5mwH0IVH7UiygDjaP1VFc

Polovinu mečeva proglasićemo klasom YES a polovinu sa NO. Samim tim za svaku kolonu računamo razlike pobednik-gubitnik. Npr razlike u broju asova i tome dodelimo klasu YES. Ako računamo razlike gubitnik-pobednik(dobijamo npr pozitivan broj duplih grešaka, negativan broj asova) dodelićemo klasi NO. Medjutim ono sto nije dobro kod ove ideje jeste da neko zaista moze da predvidi ishod meča na osnovu statistika datih kroz razlike ali mečevi su vec odigrani pa naš model nema bas smisla... Tu dolazmo na ideju broj dva koja vec govori o tome ko će pobediti u meču koji jos uvek nije odigran. Uz svakog igrača grupišemo njegove statistike. Pri susretu dva igrača njihove statistike koristimo za model. Sada imamo statistike iz prethodnih mečeva i opet pravimo razlike. Sve to pod pretpostavkom da igrači vec godinama igraju sličnim stilom. Kroz ovu ideju računali smo procente ubačenog prvog servisa,wta rangove, medjusobne susrete itd... Cilj je uneti dva igrača u model. Prelazimo na izracunavanje modela.

Klasifikacija

Koristila sam različite modele. Pušten je KNN u python-u(dobijena je tacnost od 0.69 na test skupu) kao i u spss-u.

- CRT model

Results for output field outcome

Individual Models

Comparing $R-outcome with outcome

'Partition'	1_Training		2_Testing	
Correct	1,606	82.02%	727	79.19%
Wrong	352	17.98%	191	20.81%
Total	1,958		918	

Coincidence Matrix for $R-outcome (rows show actuals)

'Partition' = 1_Training	no	yes
no	667	292
yes	60	939

'Partition' = 2_Testing	no	yes
no	317	162
yes	29	410

Performance Evaluation

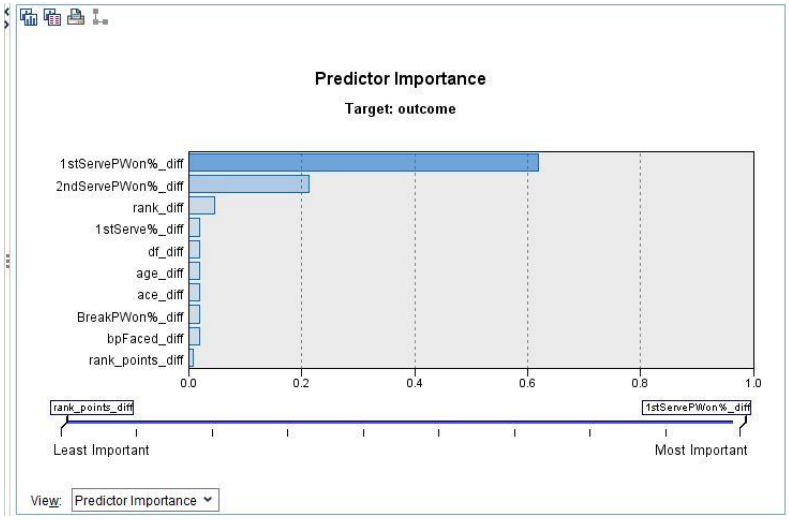
'Partition' = 1_Training	
no	0.628
yes	0.402

'Partition' = 2_Testing	
no	0.563
yes	0.405

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$R-outcome	0.942	0.884	0.923	0.845

Ovim modelom dobijamo tačnost na trening skupu od 82.02% i na test skupu od 79.19%. Sa slike mozemo pročitati i matricu konfuzije.



Uočavamo da je procenat prvog servisa jako bitan.

- SVM model uz kernel RBF

Results for output field outcome

Individual Models

Comparing $S-outcome with outcome

'Partition'	1_Training		2_Testing	
Correct	1,541	78.7%	689	75.05%
Wrong	417	21.3%	229	24.95%
Total	1,958		918	

Coincidence Matrix for $S-outcome (rows show actuals)

'Partition' = 1_Training	no	yes	$null$
no	578	42	339
yes	32	963	4
'Partition' = 2_Testing	no	yes	$null$
no	276	33	170
yes	23	413	3

Performance Evaluation

'Partition' = 1_Training	
no	0.66
yes	0.63
'Partition' = 2_Testing	
no	0.57
yes	0.661

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$S-outcome	0.991	0.982	0.982	0.964

- Logistička regresija daje

Results for output field outcome

Individual Models

Comparing $L-outcome with outcome

'Partition'	1_Training		2_Testing	
Correct	1,543	78.8%	695	75.71%
Wrong	415	21.2%	223	24.29%
Total	1,958		918	

Coincidence Matrix for $L-outcome (rows show actuals)

'Partition' = 1_Training	no	yes	$null$
no	581	39	339
yes	33	962	4
'Partition' = 2_Testing	no	yes	$null$
no	280	29	170
yes	21	415	3

Performance Evaluation

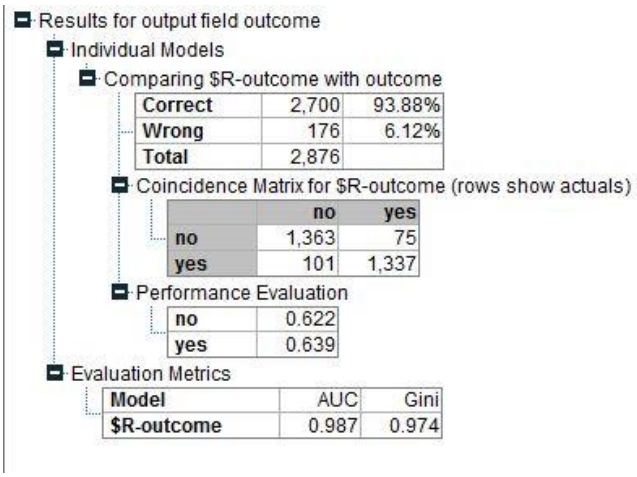
'Partition' = 1_Training	
no	0.659
yes	0.633
'Partition' = 2_Testing	
no	0.578
yes	0.67

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$L-outcome	0.991	0.982	0.982	0.965

- Kod C50 mozemo videti da je procenat prvog servisa jako bitan i da od njega zavisi ko odnosi pobedu u meču.

- CHAID



Iz kolone AUC zaključujemo da je model jako dobar posto je priblizan jedinici.

Zaključak

Zaključujemo da model sa velikom tačnošću radi. Na Kaggle mozemo naći da će ovaj data set biti obogaćen još nekim atributima poput povreda, trenutna forma itd. Samim tim to će nam zakomplikovati ali poboljšati model.

[1]