Мила Будимировић (ИН 22/2021)

Лена Рељић (ИН 21/2021)

Предикција исхода тенисних мечева на основу различитих фактора и анализа тренда 10 најбољих играча

(Тениски мечеви)

Извештај за практично истраживање

1. Увод

Предмет истраживања

Предмет истраживања је анализа и предикција исхода тениских мечева на основу различитих фактора као што су године играча, површина на којој се игра, трајање меча, ранг играча, висина играча. Ово истраживање обухвата коришћење логистичке регресије као и стабла одлучивања који могу тачно предвидети победнике мечева и турнира. Такође се ради и анализа ‘head-to-head’ резултата између играча. Истражићемо утицај појединачних фактора на резултате 10 најбољих играча.

Циљеви истраживања

1. Развити и тестирати моделе логистичке регресије и стабала одлука за предикцију победника тениских мечева и турнира.
2. Анализирати утицај различитих фактора (године, површина терена, трајање меча, ранг играча, висина) на исходе мечева.
3. Визуализовати резултате анализе кроз графичке приказе који ће показати како ови фактори утичу на учесталост победа и пораза 10 најбољих играча.
4. Израдити анализу ‘head-to-head’ мечева између врхунских играча да би се открили шаблони и трендови у њиховим резултатима.

Задаци истраживања

1. Припрема и обрада података:
    - Преузимање скупа података о АТП турнирима, играчима и тренутном рангирању
    - Чишћење и припрема података за анализу - уклањање свих недостајућих вредности из скупа података
2. Развој модела за предикцију исхода мечева:
    - Израда и тестирање логистичке регресије и стабала одлука за предикцију победника мечева на основу доступних података.
    - Валидација модела и процена тачности и прецизности предикција.
3. Спровођење ‘head-to-head’ анализе резултата.
4. Визуализација утицаја фактора на 10 најбољих тенисера.
5. Интерпретација резултата

Очекивани резултати истраживања

1. Тачност модела предикције:
    - Развијени модели логистичке регресије и стабала одлука ће показати високу тачност у предикцији победника тениских мечева и турнира. Очекује се да ће модели имати високу прецизност.

2. Анализа утицаја различитих фактора:
    - Очекујемо да ће различити фактори, као што су године играча, површина терена, трајање меча, ранг играча, и висина бити кључни као улазни подаци за модел предикције меча.

3. Визуализација података:
    - Очекујемо да појединачни фактори неће правити огроман утицај на првих 10 најбољих играча, с обзиром да је у исход меча укључено много фактора.

4. ‘Head-to-head’ анализа:
    - Анализа ‘head-to-head’ мечева ће открити специфичне факторе који доприносе доминацији или слабостима играча у одређеним дуелима.
    - Ови резултати ће пружити дубље увиде у међусобне релације врхунских играча и факторе који утичу на њихове међусобне исходе.

2. Методологија

Коришћени подаци

1. АТП турнири из више различитих година – подаци о тениским мечевима

   Садржај: идентификатори турнира (tourney_id), име турнира (tourney_name), тип подлоге (surface), величина жреба (draw_size), ниво турнира (tourney_level), датум одржавања (tourney_date), број меча (match_num), информације о победнику и губитнику меча (укључујући ИД играча, ранг, бодове, старост, доминацију руке, државу и висину), информације о самом мечу (резултат, трајање, рунде), као и статистику меча (број асова, двоструких грешака, сервиса, освојених поена).

   Обим: Овај скуп података обухвата све тениске мечеве током последњих 10 година. Имамо посебан csv фајл за сваку годину

2. АТП играчи – подаци о играчима

   Садржај: идентификатори играча (player_id), име и презиме играча (name_first, name_last), доминантна рука (hand), датум рођења (dob), националност (ioc), висину (height)

   Обим: Један csv фајл са подацима о индивидуалним играчима.

3. АТП рангирање

   Садржај: датуми (date), редослед играча на ранг листи (rank), идентификатор играча (player_id), и број бодова играча (points).

   Обим: Подаци о тренутном рангирању играча.

Сви скупови података се налазе на овом линку: https://github.com/JeffSackmann/tennis_atp

Претходна истраживања других особа над коришћеним подацима

Нисмо нашли истраживања која користе ове податке зато што су скупови података новији.

Методе истраживања

1. Припрема и обрада података
   - Преузимање података о АТП турнирима, играчима и рангирању.
   - Чишћење података: уклањање недостајућих вредности, трансформација и нормализација података.
   - Недостајуће вредности код предикције мечева и турнира смо попуњавали на другачије начине код различитих колона:

     1. За колону ‘surface’ недостајуће вредности су постављене на ‘Hard’, зато што се већина турнира игра на тврдој подлози

     2. За колоне ‘winner_hand’, ‘loser_hand’, ‘winner_ht’,’loser_ht’ као и већину колона које представљају статистику меча, недостајуће вредности смо попуњавали најчешће присутном вредношћу у одговарајућој колони. Ово смо постигли коришћењем функције mode() која враћа најчешће присутну вредност.

     3. За колоне ‘winner_rank’ и ‘loser_rank’, неодстајуће вредности смо попунили са 1000 зато што скуп података нема толико прецизне податке за мање познате и лошије рангиране играче. Играчи који имају недостајуће вредности за ранг највероватније нису најбоље рангирани.

2. Развој модела за предикцију исхода мечева

   Имамо два модела за предвиђање меча. Оба модела користе логистичку регресију. Разлика између два модела је у енкодирању ненумеричких података.

   У првом пројекту predikcija_mec1.py користимо target encoding како би могли да прикажемо које колоне највише утичу на предикцију.

   У другом пројекту predikcija_mec2.py користимо one-hot encoding за енкодирање података. Овај модел је тачнији зато што ће после енкодирања имати много више колона него први модел.

   У оба модела смо користили StandardScaler како би побољшали перформансе, нормализовали и стандардизовали податаке пре тренирања модела.

   У првом моделу смо на крају извршили анализу 10 најкориснијих колона за наш модел. Резултат смо приказали графиком.

3. Анализа head-to-head резултата

У овој анализи филтрирамо мечеве у којима је један од играча победио, а други био губитник и обрнуто. Затим бројимо победе сваког играча, као и укупан број освојених сетова и асева.

4. Графички приказ утицаја факотра на 10 најбољих тенисера
   - Филтрирање мечева где су учесници топ 10 играчи.
   - Графици који су коришћени су: графикон топлотне карте(heatmap), стубичасти (bar), пита графикон (pie), линијски графикон (lineplot)
   - Анализе које смо приказали графиконима:
     1. Утицај освојеног првог сета на исход меча
     2. Утицај броја асева на исход меча
     3. Утицај ранга играча на исход меча
     4. Утицај висине на исход меча
     5. Утицај старости играча на исход меча
     6. Број победа сваког играча на свакој подлози
     7. Број победа сваког играча на дужим мечевима

   Ова анализа пружа дубљи увид у различите аспекте игре топ 10 тенисера

5. Модел за предикцију победника турнира

За предикцију победника турнира користили смо модел стабла одлука. Предикција победника се одвија преко симулације турнира. Направили смо функцију која симулира турнир са рундама и мечевима. Шаљемо по два играча и име турнира функцији која садржи наш модел и предвиђа победника тог меча. У суштини користимо модел за предвиђање меча који се покреће толико пута колико имамо мечева. У функцији за предвиђање победника меча филтрирамо податке тако да узмемо само оне мечеве у којима је играо бар један од два играча. Делимо тежине редовима. Тежину 3 ће добити мечеви који су се одиграли на датом турниру у којима су играла обојица играча, тежину 2 ће добити мечеви између ова 2 играча али на другим турнирима и у свим осталим случајевима мечеви ће добити тежину 1. Уопштено користимо историјске податке о тениским мечевима и машинско учење за симулацију тениских турнира. Кроз процес чишћења података, припреме података за модел и симулације, програм омогућава предвиђање исхода турнира на основу специфичних критеријума.

3. Резултати

3.1. Приказ резултата

- Графици - утицај фактора на перформансе привх 10. играча

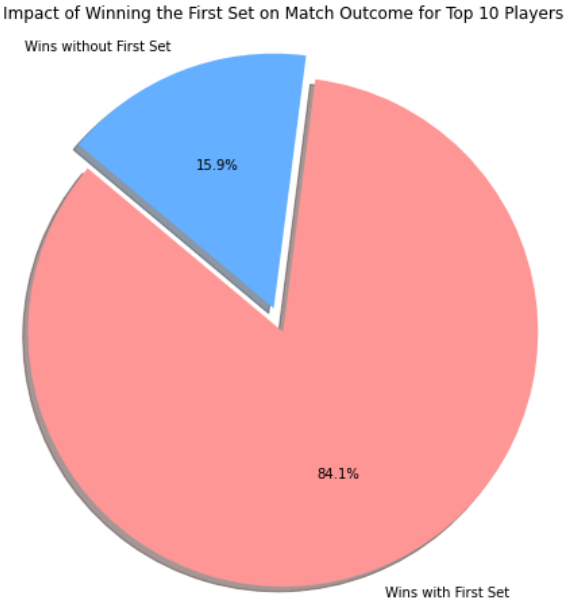


График са леве стране нам приказује утицај освојеног првог сета на исход меча. Видимо да након освојеног првог сета чак 84,1% играча заврши меч победом, док само 15,9% успе да преокрене меч након губитка првог сета. Ово нам указује на то колико је важан добар почетак у мечу.

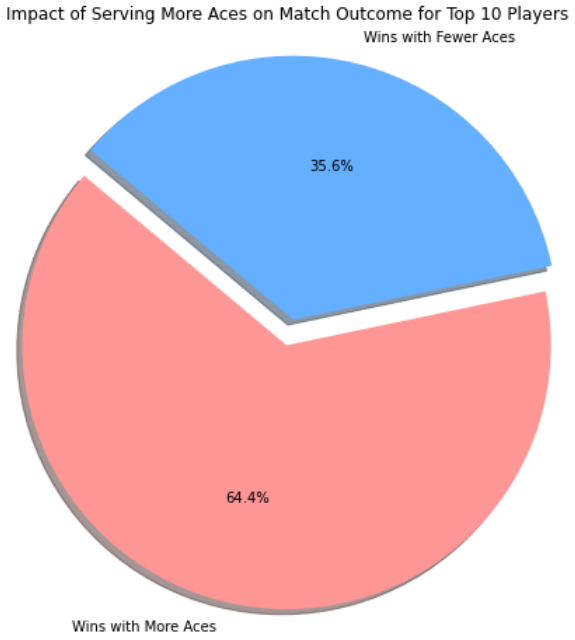
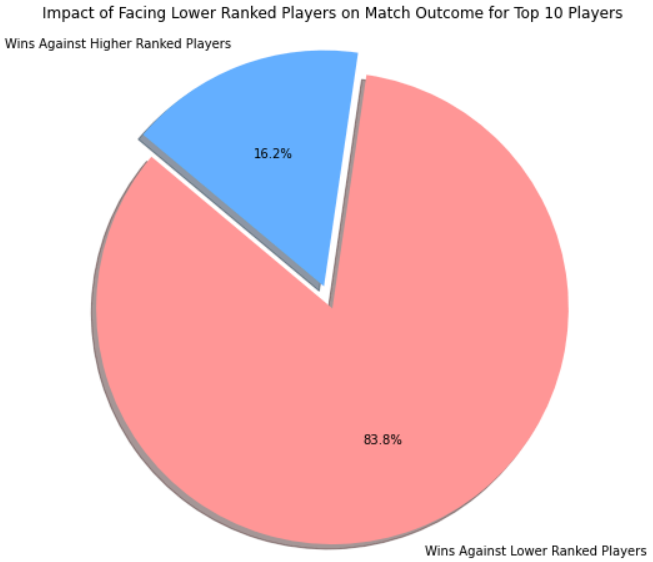
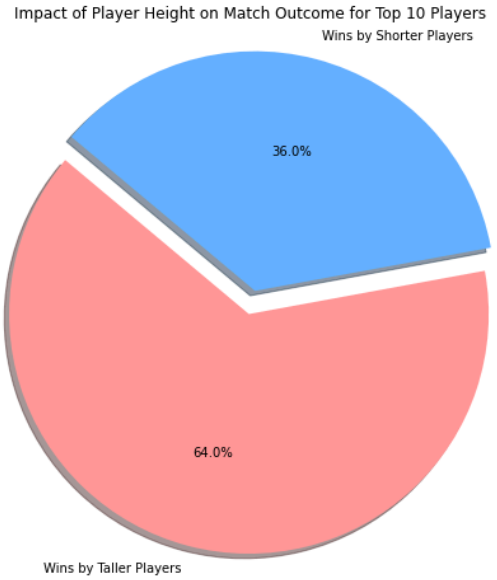


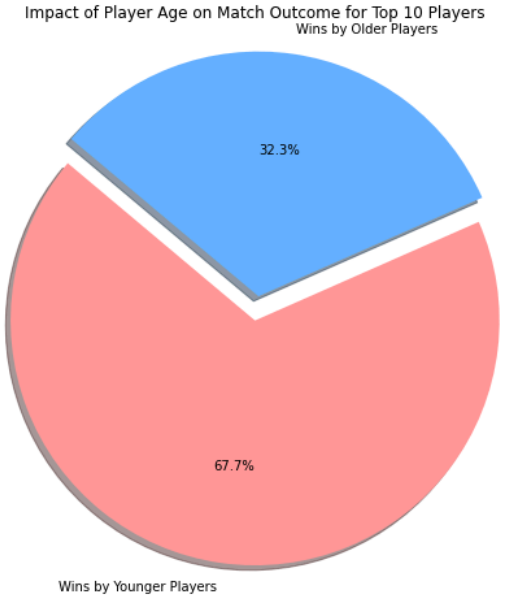
График нам указује на то колико је добар сервис битан за победу. Из приложеног видимо да тенисери који остваре већи број асева у мечу, у 64,4% случајева победе у том мечу, док ће у 35,6% случајева, тенисер који је остварио мањи број асева на крају однети победу.



Приказани граф нам одговара на питање: Да ли је ранг у тенису заиста важан? Одговор је да, као што можемо да видимо, играчи остварују победу против слабије рангираних од себе, у чак 83.8% случајева. У ретким случајевима, 16,2% играча успева да победи боље рангираног од себе.



Иако тенис не изгледа као спорт у ком је висина важан фактор, испоставило се да у 64% случајева виши играчи односе победу.



На овом графику видимо да млађи играчи ипак остварују предност над старијим. У 67,7% случајева, млађи тенисер односи победу над оним старијим.

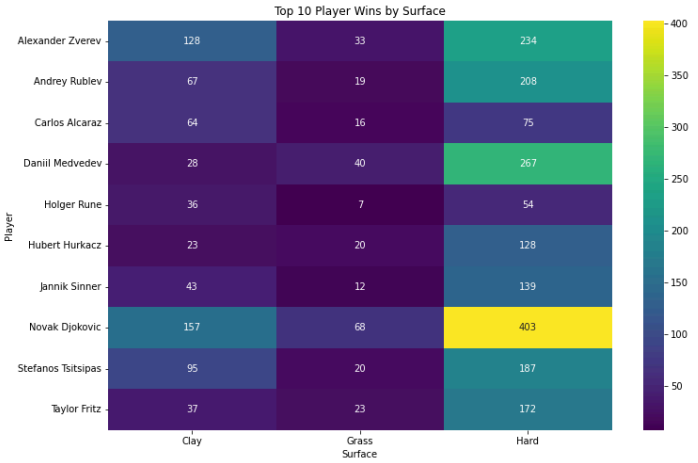


График нам приказује број победа играча на свакој подлози. Можемо закључити ком играчу која подлога највише одговара. Ђоковић постиже највише успеха на бетону.

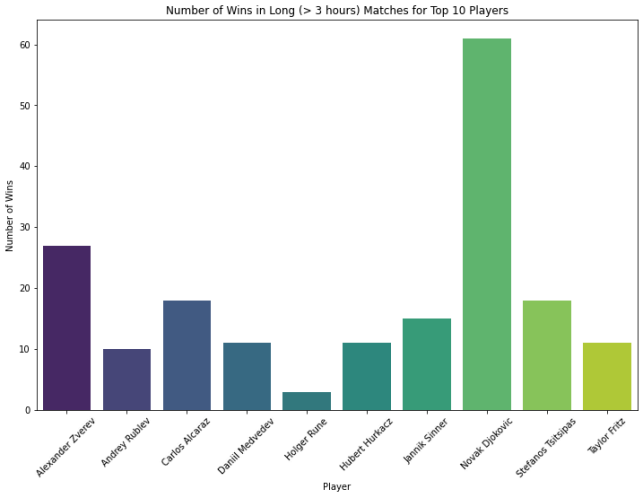


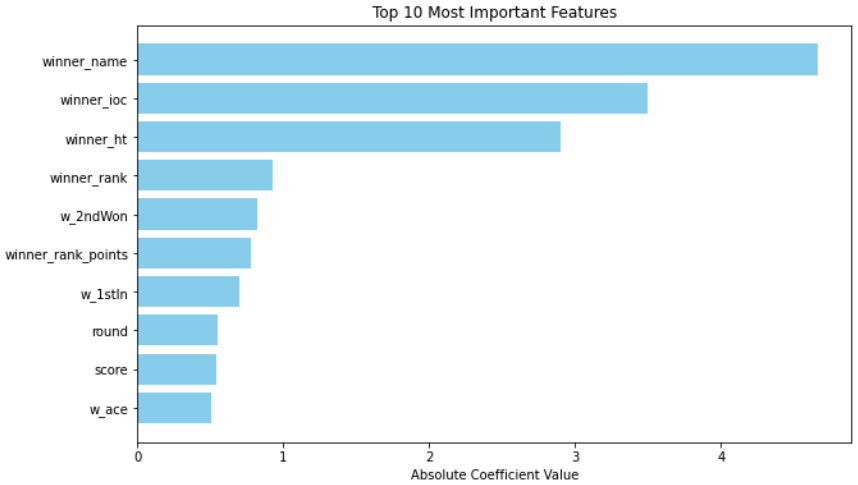
График нам приказује број победа сваког играча на мечевима дужим од 3 сата. То нам говори колико је физичка спремност сваког играча битан утицај.

- Резултат симулације турнира помоћу модела

  Као резултат симулације турнира добићемо исписан меч у терминалу

```
Simulating Australian Open with players [126774, 206173, 104745, 100644, 134770, 207989,
104925, 106421]
Round 1:
Match: Stefanos Tsitsipas vs Jannik Sinner, Winner: Stefanos Tsitsipas
Match: Rafael Nadal vs Alexander Zverev, Winner: Rafael Nadal
Match: Casper Ruud vs Carlos Alcaraz, Winner: Casper Ruud
Match: Novak Djokovic vs Daniil Medvedev, Winner: Novak Djokovic
Round 2:
Match: Stefanos Tsitsipas vs Rafael Nadal, Winner: Rafael Nadal
Match: Casper Ruud vs Novak Djokovic, Winner: Novak Djokovic
Round 3:
Match: Rafael Nadal vs Novak Djokovic, Winner: Novak Djokovic
Winner of Australian Open is Novak Djokovic
```

- Резултати утицаја фактора на предикцију победника меча



- Резултати модела

  Прецизност првог модела predikcija_mec1.py је прилично лоша на 45%.

  Док је прецизност другог модела чак 95%.

3.2. Тумачење резултата

Током анализе наших модела за предикцију мечева, установљено је да резултати нису у потпуности одговарали нашим очекивањима. Код нашег првог модела за предикцију мечева смо очекивали много већу прецизност, али због великог броја ненумеричких података и коришћења target encoding-a та прецизност је много мања.

Код другог модела велика прецизност је очекивана зато што енкодирањем података помоћу one-hot encoding-a повећавамо број колона и то за јако велик број. Ово чак може довести и до преобучавања модела што није најпожељније.

У нашем моделу, карактеристике као што су име играча, порекло, висина, ранг и рунда меча имају највећи утицај на предвиђање победника. Ови подаци се показују као најкориснији јер значајно доприносе тачности модела. Име играча и његово порекло могу указати на специфичне стилове игре и претходне перформансе, док висина играча може утицати на њихову способност сервиса и физичку доминацију. Ранг играча је директан показатељ њихове компетенције и резултата у претходним мечевима, а рунда меча може утицати на ниво притиска и умора играча. Ове карактеристике заједно пружају детаљан увид у факторе који доприносе победи у тенису.

Што се тиче фактора који утичу на 10 најбољих играча, може се закључити да 84,1% играча који освоје први сет побеђују, што истиче важност доброг почетка. Такође, играчи који остваре више асова побеђују у 64,4% случајева, наглашавајући значај доброг сервиса. Ранг је такође важан фактор, јер играчи победе слабије рангиране у 83,8% случајева. Иако висина није кључна, виши играчи побеђују у 64% мечева, док млађи играчи побеђују старије у 67,7% случајева. Ђоковић постиже највише успеха на бетону, а физичка спремност је значајна за мечеве дуже од три сата.

Код предикције турнира приметићемо важност турнира на којем играчи играју. Исти склоп играча може играти на два различита турнира, а добићемо различиту предикцију за победника. На то исто има утицај и подлога на којој се турнир игра. Наравно највећу вероватноћу да победе имају играчи који су боље рангирани као и они који имају бољи ‘headtohead’ у односу на друге играче.

4. Закључак

Анализа испуњења циљева истраживања

Један од циљева је био да развијемо и тестирамо моделе логистичке регресије и стабала одлука за предикцију победника тениских мечева и турнира, што смо успешно урадили.

Анализом утицаја различитих фактора на првих 10 играча (године, површина терена, трајање меча, ранг играча, висина играча…) дошли смо до закључка колико који фактор утиче на исход меча. Успешно смо приказали те резултате помоћу графика.

Направили смо ‘head to head’ анализу која нам пружа информацију о међусобном односу победа између два играча, као што смо навели на почетку пројекта.

Анализа остварења очекиваних резултата истраживања

Прецизност код модела није висока као што је било очекивано. За наш модел једни од најкориснијих фактора су били националност и висина, где је очекивано било да ће преовладати ранг играча.

Код визуализације није било очекивано да победа у првом сету толико повећава вероватноћу победе тог играча. Било је очекивано да ће ранг и висина доста утицати на победу и то је испунило очекивања.

Могућности за примену истраживања у пракси

Примена овог истраживања у пракси може помоћи тенисерима и њиховим тренерима, да добију увид у битне факторе који утичу на исход меча. Тренери могу скренути

пажњу играчима колико је заправо битан добар почетак меча, с обзиром да чак 84,1% играча који освоје први сет, победе у мечу. Истраживање може значити аналитичарима и свим љубитељима спорта, који желе да стекну нова сазнања и дубље разумевање учинка различитих фактора на перформансе спортиста, што им може помоћи у праћењу и предвиђању резултата у спорту.

Идеје за побољшање и разраду истраживања

Можемо побољшати наше истраживање проналаском додатних података. У скупу података које смо користиле за ово истраживање, био је одређен број недостајућих вредности, самим тим, попуњавањем недостајућих вредности се смањује прецизност модела.

Што се тиче разраде истраживања, проширили бисмо нашу анализу да обухвата и женски тенис. Ова проширена анализа би нам омогућила да упоредимо утицај различитих подлога на мушке и женске играче, што би дало додатне увиде у разлике у њиховим перформансама.

5. Литература

[1] https://www.atptour.com/en/players/atp-head-2-head

[2] https://www.atptour.com/en/rankings/singles