# *NBA Most Valuable Player Prediction using Machine Learning*

Cristian-Andrei BANTO
cristianbanto@yahoo.com

## Abstract

Premiul pentru Cel mai valoros jucător (MVP) este una dintre cele mai prestigioase distincții pe care le poate primi un jucător NBA în cariera sa. La sfârșitul fiecărui sezon regulat (de obicei în aprilie/mai), premiul MVP este înmânat unui singur jucător care este considerat demn de acest titlu. Cu toate acestea, în toate mass-media, există încă multe întrebări și dezbateri cu privire la definiția celui mai valoros jucător.

Deși s-ar putea să nu existe un răspuns cert la această discuție, instrumente precum modelele de învățare automată împreună cu date statistice relevante pot fi utilizate pentru a găsi unele modele și o perspectivă asupra logicii selecției MVP-ului NBA.

**Index Terms**: NBA, MVP, Machine-Learning.

## 1. Introducere

În acest proiect, datele istorice ale candidaților MVP din sezoanele precedente vor fi studiate. Diferite modele de învățare automată vor fi instruite și evaluate înainte de selectarea celui mai fezabil model. Apoi, modelul selectat va fi folosit pentru a face o predicție a celui mai valoros jucător al acestui sezon curent (2021-2022).

Înainte de a trece la partea de implementare, trebuie să se înțeleagă modul în care NBA decide asupra celui mai valoros jucător și cum pot fi utilizate statisticile. În prezent, alegătorii sunt 100 de membri media independenți care nu sunt afiliați la echipe și nici la jucători. Fiecare membru selectează jucători în sistemul de vot ponderat: vot pentru locul întâi (10 puncte), vot pentru locul al doilea (7 puncte), vot pentru locul trei (5 puncte), vot pentru locul al patrulea (3 puncte) și votul pentru locul cinci (1 punct). Jucătorul care primește cele mai multe „puncte” este ales cel mai valoros jucător pentru acel sezon.

Aceasta conduce la abordarea acestui experiment. Diverse statistici vor fi utilizate pentru acele „caracteristici” pe care modelul le folosește pentru a prezice MVP folosind punctele de vot MVP.

Pentru a fi consecvenți în timpul evaluării acestor numere de-a lungul timpului, punctele de vot MVP pentru fiecare jucător în particular pot fi împărțite la totalul de puncte de vot MVP pentru sezonul respectiv. Această valoare este denumită în mod obișnuit MVP Share.

MVP Share = (MVP points for particular player) / (Total MVP points)

Jucătorul care a avut cea mai mare valoare a cotei MVP câștigă premiul MVP pentru acel sezon regulat.

La început, acest exercițiu poate fi privit ca o problemă de regresie, pe care, în acest caz, variabila țintă pe care modelul de învățare automată ar încerca să o prezică este măsurarea cotei MVP. Acest lucru va fi făcut pentru fiecare sezon regulat separat, utilizând lista candidaților MVP pentru acel sezon.

Apoi, jucătorul (pentru un anumit sezon) care a avut cea mai mare valoare estimată a cotei MVP este etichetat ca MVP prezis de către model. Numele MVP-ului real în timpul sezonului respectiv este comparat cu MVP-ul prezis pentru a verifica dacă modelul a fost corect sau incorect.

## 2. Implementare

Pentru a evalua un model pentru toate sezoanele (unde sunt disponibile date relevante), poate fi folosită o abordare inspirată de „shuffling”. De exemplu, pentru fiecare sezon terminat (din 1980 până în 2021), modelul trebuie testat pentru datele unui anumit sezon și antrenat pe datele altor sezoane care nu au fost selectate. Acest lucru se va repeta până când datele pentru toate sezoanele individuale vor fi testate și prezise pentru a calcula acuratețea generală.

Un exemplu ar fi urmatorul caz: o iteratie, unde avem ca date de antrenare liste cu candidatii pentru MVP din 1980-1992 si 1994-2021 iar ca date de test o sa luam lista cu candidatii din anul 1993, unde modelul o sa incerce sa prezica MVP Share-ul.

Deoarece este cunoscută cota MVP reală pentru 1993, aceasta poate fi utilizată pentru a compara cu cotele MVP prezise. În general, MAE (Eroarea medie absolută) și $R^2$ (coeficientul de determinare) sunt câteva metrici populare pentru evaluarea modelelor de regresie. O valoare mai mică a MAE (intervalul de la 0 la ∞) și o valoare mai mare a lui $R^2$ (intervalul de la 0 la 1) este de dorit pentru un model performant.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

$x_i$ = actual value
$y_i$ = predictions
$n$ = sample size

$$R^2 = 1 - \frac{RSS}{TSS}$$

$R^2$ = coefficient of determination
$RSS$ = sum of squares of residuals
$TSS$ = total sum of squares

Deși în acest context, aceste două metrici nu sunt singurii factori importanți de luat în considerare. De exemplu, atunci când se testează un sezon în care cursa MVP a fost foarte apropiată între primii doi candidați, cotele MVP estimate pot fi foarte apropiate de cotele MVP reale pentru acești doi candidați. Cu toate acestea, candidatul care a primit cea mai mare cotă de MVP estimată a modelului poate să nu fi primit cea mai mare cotă de MVP în cursa reală.

Am creat o eticheta numita Label unde o sa vedem daca modelul a prezis correct sau nu MVP-ul din anul respectiv, de care o sa ne folosim si pentru calculul acuratetii:

Accuracy = (correct labels) / (correct labels + incorrect labels)

## 2.1. Vizualizarea si interpretare datelor de antrenare

Sunt necesare două seturi de date: primul contine statistici istorice ale candidaților MVP, inclusiv toți câștigătorii și învinșii care au fost luați în considerare și al doilea cu statisticile candidaților MVP în sezonul 2021-2022.

Trebuie observate relatiile dintre caracteristici și variabila de răspuns, în cazul nostru NBA Share.

Mai jos, unele caracteristici generale au fost reprezentate în raport cu cota MVP. Numărul de victorii ale echipelor și clasarea echipelor sunt de obicei câțiva factori cruciali care determină discuția MVP. De asemenea, a fost trasată o statistică avansată numită VORP (Valoare peste jucător de înlocuire: folosită pentru a măsura contribuția generală a unui jucător la echipă).

Punctele de dispersie reprezintă fiecare candidați din perioada 1980-2021, cu câștigătorii MVP actuali evidențiați cu albastru. In imaginea de mai jos avem in grafic pe axa x numarul de victorii din sezon in functie de cota MVP (share).

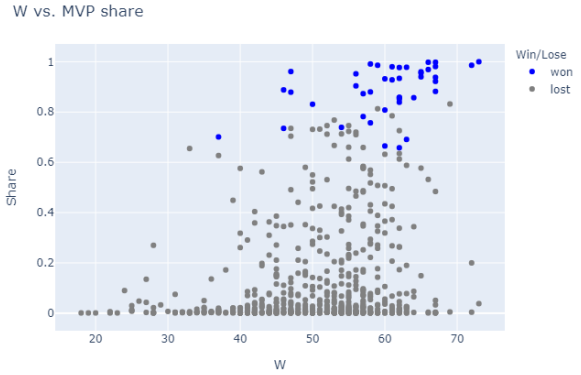


Figura 1. Relația dintre numărul de victorii și cota MVP

O altă caracteristică interesantă a fost rata de utilizare. Rata de utilizare este o estimare a procentului de joc în echipă în care a fost implicat un jucător.

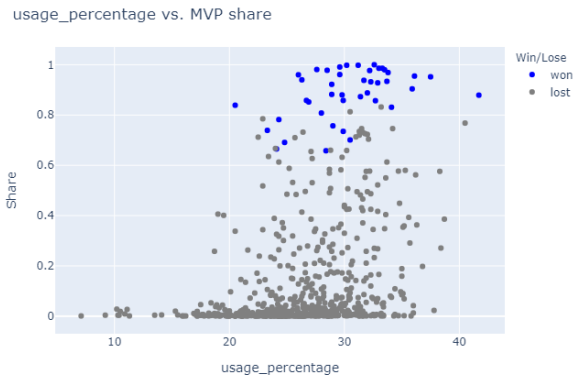


Figura 2. Relația dintre usage_percentage și cota MVP

În acest moment, aceste variabile pot părea că au o relație oarecum liniară cu valoarea cotei MVP. Ar putea fi valabil să le considerăm caracteristici model pe măsură ce experimentele sunt efectuate.

Tabelul principal avea la un moment dat până la 45 de coloane (inclusiv informații de bază, cum ar fi numele jucătorilor și ale echipelor), dar numai câteva coloane au servit ca caracteristici utile. Utilizarea numai a caracteristicilor relevante ar ajuta la reducerea zgomotului și la creșterea performanței modelelor de învățare automată.

O modalitate de a identifica acest lucru a fost printr-o metodă numită Mutual information. Mutual information este o funcție care măsoară asocierea dintre caracteristică și țintă, care poate părea similară cu corelația care detectează relații liniare. Cu toate acestea, această măsurătoare este capabilă să detecteze orice fel de relație (și nu doar liniară).

Mai jos avem scorul pentru Informatia mutuala:

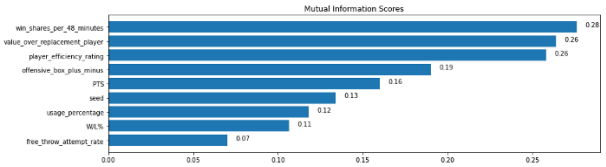


Figura 3. Mutual information – scorul pentru caracteristici

In figura de mai sus avem scorurile finale pentru informatia mutuala, practic avem cele mai importante caracteristici care sunt folosite pentru a calcula cota MVP, si a inmana titlul de MVP. Putem observa ca mai multe caracteristici cum ar fi: numarul pase decisive sau recuperari nu se afla in topul celor mai importante caracteristici necesare pentru a castiga titlul de MVP.

Pentru modelarea datelor au fost selectate mai multe modele de regresie:

- Linear Regression
- Random Forest Regressor
- XGBoost Regressor (eXtreme Gradient Boosting)
- LightGBM Regressor (Light Gradient Boosting Machine)

Regresia liniară a fost selectată deoarece multe caracteristici păreau să aibă o relație liniară cu variabila răspuns. Au fost utilizate trei modele bazate pe arbore (Random Forest, XGBoost, LightGBM). Random Forest este o metodă de învățare prin ansamblu care funcționează prin construirea unei multitudini de arbori de decizie. XGBoost și LightGBM sunt câteva variante mai eficiente și mai puternice ale modelelor de învățare în arbore. Pentru aceste modele bazate pe arbore, au fost efectuate diferite eforturi de reglare a parametrilor pentru a găsi cei mai buni parametri care ar produce cea mai înaltă performanță.

Pentru a putea face o prezicere pentru anul 2022, trebuie proiectata o noua variabila, Value over replacement player, deoarece sezonul nu este gata, iar aceste date nu exista in intregime. Pentru a proiecta acest „VORP ajustat”, vor fi necesare statisticile VORP curente și informațiile despre numărul de jocuri.

Adjusted VORP = ((current VORP) / (games played) * games left) + (current VORP)

Această metodă nu este nicidecum convențională sau ideală, dar presupune VORP-ul unui jucător până la sfârșitul sezonului (presupunând că acesta joacă restul sezonului la o performanță statistică consistentă). Încă o dată, probabil că acesta nu va fi foarte aproape de VORP real la sfârșitul sezonului, dar VORP-ul fiecărui jucător din 2022 va fi transformat în același mod pentru consecvență.

## 3. Rezultate

| Model Performance Summary | | | |
|---|---|---|---|
| Model | Average MAE | Average R squared | Accuracy |
| Linear Regression | 0.145281 | 0.471695 | 0.642857 |
| Random Forest Regressor | 0.104024 | 0.607992 | 0.738095 |
| XGBoost Regressor | **0.103273** | 0.606266 | **0.833333** |
| LGBM Regressor | 0.106055 | **0.611992** | 0.761905 |

Figura 4. Performanțele modelelor

Dintre cele patru modele, modelul XGBoost a avut cea mai bună performanță, cu o precizie de 83,33% (a prezis corect 35 de sezoane din 42 de sezoane totale) și cea mai scăzută medie MAE. Modelul LGBM a avut cel mai mare $R^2$ mediu, dar a prezis gresit încă trei sezoane (32 corecte din 42 de sezoane totale). Modelul liniar a avut cele mai slabe rezultate pe toate cele trei categorii și va fi eliminat din analiză.

Pentru restul acestei analize, rezultatele modelului XGBoost vor fi examinate puțin mai detaliat în comparație cu celelalte modele, începând cu importanța caracteristicilor sale.

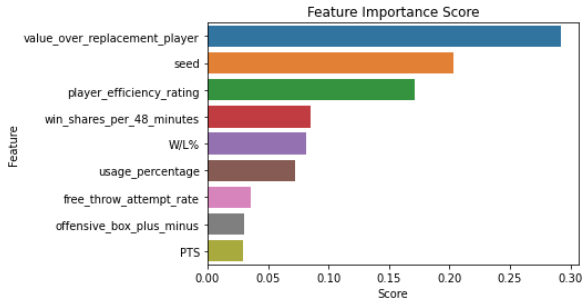


Figura 5. Importanța caracteristicilor (XGBoost)

Așa cum era de așteptat, statisticile avansate, cum ar fi VORP, cotele de câștig la 48 de minute și evaluarea eficienței jucătorilor, precum și statisticile echipei, cum ar fi locul ocupat (seed) și procentajul de victorii si infrangeri (W/L%), au un punctaj ridicat în ceea ce privește importanța caracteristicii. Singura statistică de scor general rămasă a fost punctele per joc (PTS), care încă a contribuit puțin la algoritm. Modelul s-a bazat în mare măsură pe statisticile privind eficiența individuală, împreună cu valorile performanței echipei. Deși procentul de utilizare nu evidențiază eficiența jucătorului, acesta poate sublinia dependența unei echipe de un jucător pe tot parcursul sezonului.

## 4. Interpretarea rezultatelor

Pentru validarea modelelor, cursa MVP din diferiți ani va fi inspectată atât pentru cazul în care modelul a funcționat bine, cât și când nu a funcționat.
În urmatoarea imagine avem rezultatele prezise de modelul XGBoost din ultimii 13 ani.
După cum sa menționat mai devreme, conform rezumatului modelului XGBoost, 35 din totalul de 42 de MVP-uri ale sezoanelor au fost ghicit corect cu o precizie de 83,33%. Chiar daca modelul a prezis corect ultimii 13 ani, exista anumiti ani in care pot sa existe anumite discutii, cum ar fi 2011, cand Derrick Rose a fost MVP, respectiv 2017, cand titlul a fost castigat de Russell Westbrook.

| | year | MAE | R squared | Predicted MVP | Actual MVP | Label |
|---|---|---|---|---|---|---|
| **29** | 2009 | 0.131786 | 0.586624 | LeBron James | LeBron James | correct |
| **30** | 2010 | 0.124349 | 0.576489 | LeBron James | LeBron James | correct |
| **31** | 2011 | 0.105977 | 0.688991 | Derrick Rose | Derrick Rose | correct |
| **32** | 2012 | 0.119754 | 0.582999 | LeBron James | LeBron James | correct |
| **33** | 2013 | 0.069011 | 0.848391 | LeBron James | LeBron James | correct |
| **34** | 2014 | 0.099713 | 0.719490 | Kevin Durant | Kevin Durant | correct |
| **35** | 2015 | 0.077365 | 0.861898 | Stephen Curry | Stephen Curry | correct |
| **36** | 2016 | 0.123857 | 0.672600 | Stephen Curry | Stephen Curry | correct |
| **37** | 2017 | 0.263471 | 0.011375 | Russell Westbrook | Russell Westbrook | correct |
| **38** | 2018 | 0.080505 | 0.876323 | James Harden | James Harden | correct |
| **39** | 2019 | 0.157135 | 0.635266 | Giannis Antetokounmpo | Giannis Antetokounmpo | correct |
| **40** | 2020 | 0.096595 | 0.846305 | Giannis Antetokounmpo | Giannis Antetokounmpo | correct |
| **41** | 2021 | 0.109079 | 0.734734 | Nikola Jokić | Nikola Jokić | correct |

Figura 6. Preziceri corecte 2009-2021

### 4.1 2011 MVP Race

| | Player | seed | W/L% | WS/48 | usage_percentage | player_efficiency_rating | value_over_replacement_player | Share | predicted_share |
|---|---|---|---|---|---|---|---|---|---|
| | Derrick Rose | 1.0 | 0.756 | 0.208 | 0.322 | 23.5 | 6.7 | 0.977 | 0.769880 |
| | Dwight Howard | 4.0 | 0.634 | 0.235 | 0.272 | 26.1 | 5.4 | 0.531 | 0.263934 |
| | LeBron James | 2.0 | 0.707 | 0.244 | 0.315 | 27.3 | 7.8 | 0.431 | 0.624091 |

Figura 7. Clasamentul pentru MVP 2011

Pentru anul 2011, mulți susțin că LeBron James sau chiar Dwight Howard au meritat premiul în fața lui Derrick Rose. De fapt, LeBron James l-a depășit pe Rose în aproape toate scorurile de box și statisticile avansate, susținând acest caz. Cu toate acestea, Derrick Rose a avut o rată de utilizare puțin mai mare și un record de echipă, ajutându-l să-l devanseze pe LeBron.

Acest lucru poate fi observat și printr-o vizualizare a valorilor SHAP, care este utilă atunci când înțelegeți modul în care un model complex prezice rezultatul. Valorile SHAP măsoară impactul fiecărei caracteristici în procesul de luare a deciziilor.

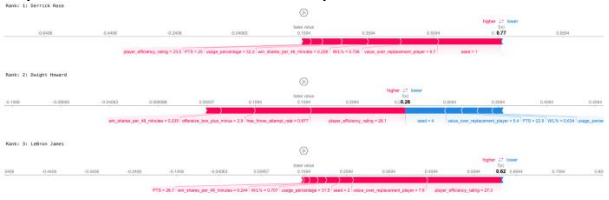

Figura 8. Valorile SHAP pentru cursa MVP 2011

Barele roșii indică caracteristici care ajută fiecare candidat în cazul prezicerii cotelor MVP mari. În schimb, barele albastre indică exact contrariul sau caracteristicile care le-au afectat. Amploarea impactului acestui lucru asupra cotelor MVP poate

fi observată prin lungimea segmentului de bară al fiecărei caracteristici.

Pentru Derrick Rose și LeBron James, nu există caracteristici ale modelului care să le strice șansele. Cu toate acestea, se pare că a fi în echipa cap de serie numărul 1 a făcut o mare diferență pentru Rose. Pentru Dwight Howard, care a ajuns pe locul 2 în viața reală, fiind într-un cap de serie mult mai scăzut (al 4-lea) și alte câteva statistici i-au scăzut valoarea.

### 4.2 2017 MVP Race

| Player | seed | W/L% | WS/48 | usage_percentage | player_efficiency_rating | value_over_replacement_player | Share | predicted_share |
|---|---|---|---|---|---|---|---|---|
| Russell Westbrook | 6.0 | 0.573 | 0.224 | 0.417 | 30.6 | 9.3 | 0.879 | 0.627730 |
| James Harden | 3.0 | 0.671 | 0.245 | 0.342 | 27.4 | 8.0 | 0.746 | 0.527090 |
| Kawhi Leonard | 2.0 | 0.744 | 0.264 | 0.311 | 27.6 | 7.1 | 0.495 | 0.467705 |

Figura 9. Cursa pentru MVP din 2017

2017 a fost o cursă dezordonată, valoarea $R^2$ fiind foarte scăzută, la 0,01. Din 2000, Westbrook este primul MVP care face parte din echipa care a terminat pe locul 6 sezonul regulat. În ciuda acestui fapt, modelul a prezis utilizând corect evaluarea eficienței jucătorului, VORP și metrica ratei de utilizare pe care Westbrook a dominat (41,65%, care este recordul unui singur sezon). Între timp, modelul a prezis cote mari de MVP pentru Kevin Durant și Stephen Curry ceea ce a afectat MAE și $R^2$.

Trecem la partea in care modelul a prezis gresit:

| | year | MAE | R squared | Predicted MVP | Actual MVP | Label |
|---|---|---|---|---|---|---|
| 2 | 1982 | 0.079323 | 0.264734 | Magic Johnson | Moses Malone | incorrect |
| 13 | 1993 | 0.128876 | 0.593915 | Michael Jordan | Charles Barkley | incorrect |
| 14 | 1994 | 0.084023 | 0.611711 | David Robinson | Hakeem Olajuwon | incorrect |
| 21 | 2001 | 0.135485 | 0.476847 | Shaquille O'Neal | Allen Iverson | incorrect |
| 25 | 2005 | 0.181874 | -0.113709 | Dwyane Wade | Steve Nash | incorrect |
| 26 | 2006 | 0.192149 | -0.010072 | Dirk Nowitzki | Steve Nash | incorrect |
| 28 | 2008 | 0.110665 | 0.622425 | LeBron James | Kobe Bryant | incorrect |

Figura 10. Preziceri incorecte

Cel mai mare accident a avut loc în anii 2005 și 2006, când Steve Nash a câștigat premii consecutive. Modelul nu numai că a ghicit jucătorii greșiți, ci MAE este mare și $R^2$ este în negativ. Cursa MVP pentru acești doi ani a fost un dezastru absolut pentru modelul de prezis după antrenamentul din ceilalți ani.

mvp_race_2005

| | Player | seed | W/L% | Share | predicted_share |
|---|---|---|---|---|---|
| 450 | Steve Nash | 1.0 | 0.756 | 0.839 | 0.009219 |
| 451 | Shaquille O'Neal | 1.0 | 0.720 | 0.813 | 0.242545 |
| 452 | Dirk Nowitzki | 3.0 | 0.707 | 0.275 | 0.161279 |
| 453 | Tim Duncan | 2.0 | 0.720 | 0.258 | 0.248626 |
| 454 | Allen Iverson | 7.0 | 0.524 | 0.189 | 0.049162 |
| 455 | LeBron James | 8.5 | 0.512 | 0.073 | 0.155844 |
| 456 | Tracy McGrady | 5.0 | 0.622 | 0.035 | 0.114428 |
| 457 | Dwyane Wade | 1.0 | 0.720 | 0.034 | 0.293732 |
| 458 | Amar'e Stoudemire | 1.0 | 0.756 | 0.032 | 0.207200 |
| 459 | Ray Allen | 4.0 | 0.634 | 0.032 | 0.053268 |
| 460 | Kevin Garnett | 9.0 | 0.537 | 0.012 | 0.214726 |
| 461 | Gilbert Arenas | 4.5 | 0.549 | 0.003 | 0.088578 |
| 464 | Shawn Marion | 1.0 | 0.756 | 0.001 | 0.147032 |
| 462 | P.J. Brown | 15.0 | 0.220 | 0.001 | 0.007065 |
| 463 | Marcus Camby | 7.0 | 0.598 | 0.001 | 0.007065 |

mvp_race_2006

| | Player | seed | W/L% | Share | predicted_share |
|---|---|---|---|---|---|
| 465 | Steve Nash | 3.0 | 0.659 | 0.739 | 0.110369 |
| 466 | LeBron James | 3.0 | 0.610 | 0.550 | 0.567135 |
| 467 | Dirk Nowitzki | 2.0 | 0.732 | 0.435 | 0.623741 |
| 468 | Kobe Bryant | 6.0 | 0.549 | 0.386 | 0.503601 |
| 469 | Chauncey Billups | 1.0 | 0.780 | 0.344 | 0.217181 |
| 470 | Dwyane Wade | 2.0 | 0.634 | 0.070 | 0.443376 |
| 471 | Elton Brand | 5.0 | 0.573 | 0.040 | 0.151996 |
| 472 | Tim Duncan | 1.0 | 0.768 | 0.026 | 0.175536 |
| 473 | Tony Parker | 1.0 | 0.768 | 0.007 | 0.035607 |
| 475 | Shawn Marion | 3.0 | 0.659 | 0.001 | 0.199527 |
| 474 | Allen Iverson | 9.0 | 0.463 | 0.001 | 0.173668 |

Figura 11. Cursa pentru MVP 2005&2006

Privind cursa în detaliu, modelul a prezis cote de MVP foarte scăzute pentru Steve Nash. Nu numai asta, modelul de predicție a favorizat alți candidați precum Dwyane Wade (2005) și Dirk Nowitzki (2006). Prin mass-media, s-a spus că premiile pentru ambii ani au fost foarte discutabile, oamenii susținând că Shaquille O'Neal (2005) și Kobe Bryant (2006) au meritat premiul în fața lui Nash. Acesta poate fi unul dintre câteva cazuri în care statisticile nu spun întreaga poveste, deoarece mulți fani știu impactul intangibil pe care l-a avut Nash asupra sistemului ofensiv al echipei sale.

### 4.3 2022 MVP Race

Pentru cursa pentru MVP 2022, cele mai bune 3 modele (XGboost, LightGBM, Random Forest) au fost instruite pe date istorice (din 1980-2021) și utilizate pentru a prezice datele candidaților MVP 2021-2022. Rezultatele au fost următoarele:

| | XGBoost | LightGBM | Random Forest |
|---|---|---|---|
| 1st Place | Nikola Jokic | Nikola Jokic | Nikola Jokic |
| Predicted MVP share | 0.679943 | 0.519327 | 0.533504 |
| 2nd Place | Giannis Antetokounmpo | Giannis Antetokounmpo | Giannis Antetokounmpo |
| Predicted MVP share | 0.565778 | 0.512879 | 0.49226 |
| 3rd Place | Joel Embiid | Joel Embiid | Joel Embiid |
| Predicted MVP share | 0.425511 | 0.329178 | 0.405678 |

Figura 12. Cursa pentru MVP din 2022

Interesant este că toate cele trei modele au prezis că Nikola Jokić va fi MVP-ul din 2022.

Cota estimată de MVP a fost mare pentru Jokić (cota estimată = 0,68) pentru că a excelat în statistici avansate, în ciuda faptului că echipa sa se afla pe locul 6.

Giannis (cota estimată = 0,57) a fost foarte aproape în cursă, dar a avut statistici puțin mai puțin avansate în comparație cu Jokić.

Embiid (cota estimată = 0,43) care a condus cursa în lunile februarie și martie, a căzut în clasament cel mai probabil din cauza accidentării sale, ceea ce a făcut ca echipa sa să cadă pe locul 4.

Acest lucru arată că este foarte dificil pentru model să prezică cu exactitate MVP-ul în timpul sezonului live. Cu toate acestea, se pare că modelul nu face o treabă atât de proastă când datele sezonului regulat sunt finalizate.

## 5. Concluzii

A face predicții în sport este o sarcină foarte dificilă și este întunecată de multă incertitudine. În acest experiment, au fost construite și analizate modele de predicție pentru a prezice premiul MVP al NBA. Deși cifrele cu siguranță nu spun întreaga poveste, au existat câteva modele interesante de metrici și statistici care au contribuit la modele. Până la urmă, cu aceste modele s-a făcut prognoza MVP-ului 2022. În viitor, va fi foarte interesant de văzut dacă modelele pot efectua și ajusta previziunile pentru a se apropia de opiniile presei și ale experților în domeniu.

# 6. Referințe


[1] Hashmi, F. (2021, November 27). Data Science interview questions for IT industry part-3: Supervised ML. Thinking Neuron. Retrieved February 4, 2023, from https://thinkingneuron.com/data-science-interview-questions-for-it-industry-part-3-supervised-ml/#XGBoost.

[2] https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.

[3] Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010 Nov;107(44):776-82. doi: 10.3238/arztebl.2010.0776. Epub 2010 Nov 5. PMID: 21116397; PMCID: PMC2992018.