Téma címe: Játékosérték-becslése adatelemzési módszerekkel

Feladat

A projekt során először önálló munka keretében kerül sor a célkitűzések pontosítására és a releváns adatforrások felkutatására. Ezt követi az adatok beszerzése és előkészítése, amely magában foglalja az adatgyűjtést, tisztítást, normalizálást, valamint új mutatók kiszámítását és kategóriák kezelését. A modellépítés szakaszában különböző statisztikai és gépi tanulási regressziós modellek (pl. lineáris regresszió, Random Forest, GBM) kerülnek kipróbálásra, melyek teljesítményét MAE, MAPE és MSE mutatók alapján értékeljük. Az eredményeket kiértékeljük és vizualizáljuk, majd kiválasztjuk és finomhangoljuk a legjobb modellt. A projekt írásbeli beszámolóval és szóbeli prezentációval zárul, melyek elkészítése és javítása a félév utolsó szakaszában történik, a tanulságok összegzésével együtt.

1. A laboratóriumi munka környezetének ismertetése, a munka előzményei és kiindulási állapota

1.1 Bevezető

A futball világában az átigazolási díjak folyamatosan emelkednek, és az egyes játékosok piaci értékének megbízható becslése mind klubszinten, mind elemzői oldalon kiemelten fontos. A becslés mögött üzleti, sportstratégiai és adatelemzési szempontok is húzódnak: a kluboknak egyre inkább adatalapú döntéseket kell hozniuk, amikor játékosokat vásárolnak vagy értékesítenek.

A jelen projekt célja egy prediktív modell létrehozása, amely a játékosokra és az átigazolási környezetre vonatkozó különböző attribútumok alapján, a data science eszköztárát felhasználva előrejelzi a becsült átigazolási díjat. A projekt során nyílt forrású adatokat (pl. Transfermarkt) használtunk, amelyeket több lépésben tisztítottunk, átalakítottunk és kiegészítettünk.

Ez a munka egy különálló, önálló féléves projektként készült, nem kapcsolódik kari vagy tanszéki nagyobb kutatási programhoz. A feladat önálló problémamegfogalmazással indult, a teljes adat-előkészítés, modellezés, értékelés és dokumentáció egyéni munkában valósult meg.

1.2 Elméleti összefoglaló

A digitális adatok gyors növekedése és az adattárolási, valamint -feldolgozási kapacitások fejlődése lehetővé tették, hogy az adatelemzés (data analysis) mára az ipar, a tudomány és a társadalom szinte minden területén nélkülözhetetlenné váljon. Az üzleti életben, az egészségügyben, a közlekedésben és a sportban is egyre nagyobb igény mutatkozik arra, hogy a döntéshozatal és a stratégiaalkotás adatalapon történjen. Az adatelemzés egyik kiemelt ága a gépi tanulás (machine learning) [1], amely algoritmusok segítségével képes a meglévő adatokból tanulni, és azokra építve előrejelzéseket, becsléseket készíteni.

A gépi tanulás eszköztára rendkívül sokrétű. A felhasználás módja szerint megkülönböztethetünk például osztályozási (klasszifikációs) feladatokat, amelyeknél kategóriákba kell sorolni egy-egy adatpontot, illetve regressziós feladatokat, amelyek során folytonos értéket próbálunk becsülni egy vagy több bemeneti változó alapján. Az átigazolási díjak előrejelzése ez utóbbi kategóriába esik, hiszen a célváltozó – a játékos várható ára – egy

numerikus érték, amelyet minél pontosabban szeretnénk meghatározni.

A regresszióhoz számos módszer áll rendelkezésre. A legegyszerűbb a lineáris regresszió [2], amely azt feltételezi, hogy a bemeneti változók és a célváltozó között lineáris kapcsolat áll fenn. Bár ez sok esetben korlátozó, a lineáris modell nagy előnye az interpretálhatóság, vagyis, hogy az egyes jellemzők hatása jól követhető. A komplexebb modellek közé tartoznak az ensemble-alapú tanulóalgoritmusok, mint például a Random Forest [2] és a Gradient Boosting [2] gépi tanulási modellek. Ezek több döntési fából állnak, amelyek kombinált működése lehetővé teszi a nemlineáris összefüggések felismerését, és jellemzően jobb predikciós teljesítményt nyújtanak, különösen heterogén adatok esetén.

Az átigazolási piac sajátosságai – például a játékosok életkora, pozíciója, nemzetisége, a ligák különböző erősségei vagy a klubváltások iránya – nagymértékben befolyásolják egy játékos értékét. Az adatok elemzése során gyakran előfordulnak kiugró értékek (outlierek), például világsztárok extrém magas ára. Az ilyen értékek torzíthatják a modell működését, ezért célszerű ezek kezelésére, például trimmelt eloszlásokra vagy logaritmikus skálázásra támaszkodni.

A gépi tanulási modellek értékelése szintén fontos része a predikciós rendszerek fejlesztésének. A regressziós modellek jellemző hibamértékei közé tartozik a MAE (mean absolute error) [2], amely az átlagos abszolút eltérést méri; az MSE (mean squared error) [2], amely a nagyobb hibákat erősebben bünteti, valamint a MAPE (mean absolute percentage error) [2], amely az eltéréseket százalékos formában adja meg. Ezek segítségével összehasonlíthatók különböző modellek teljesítményei, és kiválasztható a legjobban illeszkedő megoldás.

A futballista-átigazolási piac értékelése és modellezése tehát nem csupán üzleti vagy sportgazdasági szempontból releváns, hanem kiváló példa arra is, hogyan alkalmazhatók korszerű adatfeldolgozási és gépi tanulási módszerek egy összetett, sokténye zős valós probléma elemzésére.

1.3 A munka állapota, készültségi foka a félév elején

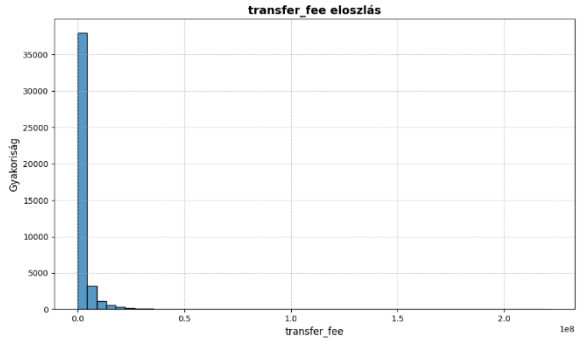
A félév kezdetén kifejezetten ezzel a témával, vagyis a futballista-átigazolások gépi tanulással történő modellezésével még nem foglalkoztam, azonban a data science területével korábban már több projekt keretében is volt alkalmam megismerkedni. Rendelkeztem tehát az alapvető módszertani ismeretekkel az adattisztítás, jellemzőképzés, valamint a regressziós modellezés terén. A konzulensemtől egy nagyméretű .csv formátumú adatállományt kaptam kézhez, amely a játékosokra, klubokra, bajnokságokra és átigazolásokra vonatkozó részletes információkat tartalmazott. Ez szolgált a féléves munka alapjául, és ebből kiindulva kezdtem meg az adat előkészítését, az elemzés strukturálását, valamint a predikciós modell megtervezését és kiépítését.

2. Az elvégzett munka és eredmények ismertetése

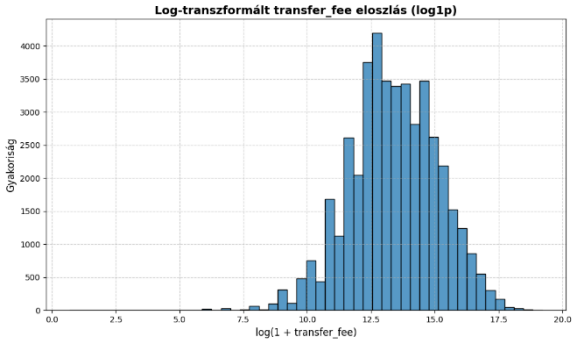
2.1 Adatbetöltés és kezdeti vizsgálat

A munka első lépése a konzulensemtől kapott nyers adathalmaz beolvasása [1] volt, amely egy .csv formátumú táblázatban tartalmazta a játékosok átigazolásaival kapcsolatos információkat: többek között a játékosok nevét, életkorát, nemzetiségét, pozícióját, az átadó és fogadó klub nevét, a ligák szintjét, valamint az átigazolás díját. Az adathalmaz több mint egymillió sort tartalmazott, így már az elején szükség volt hatékony szűrésre és letisztításra, hogy az elemzés elvégezhető legyen.

Az adatok betöltése és a duplikált vagy irreleváns rekordok kiszűrése után megvizsgáltam, hogy mely mezők tartalmaznak hiányzó értékeket, és azokat külön kezeltem (pl. NaN [not a number] kiszűrése, értelmes default értékek megadása, típuskonverziók).



1: ábra. Az átigazolási díjak eloszlása az eredeti skálán



2: ábra. Az átigazolási díjak logaritmizált eloszlása

2.2 Adattisztítás és szűrés [1]

A kezdeti fázist követően az adathalmazt szűkítettem: eltávolítottam a NaN értékeket tartalmazó rekordokat a transfer_fee mezőből, majd csak azokat a megfigyeléseket tartottam meg, amelyeknél az átigazolási díj pozitív érték volt. Emellett eltávolítottam azokat az oszlopokat is, amelyek vagy irrelevánsak voltak, vagy túl nagy arányban tartalmaztak hiányzó adatot. Így az 1 millió soros adathalmazunkból viszonylag gyorsan 50 ezer sor maradt csupán.

A minutes_played mezőnél például kiszűrtem az irreálisan magas, vélhetően elírt értékeket (pl. >7000 perc), és néhány oszlopnál az adatokat hisztogramok alapján vizuálisan is ellenőriztem, hogy kiszűrjem a hibás vagy kiugró értékeket.

Több olyan változót is tartalmazott az adathalmaz, amelyek szöveges, bináris jelentésű információt hordoztak (pl. is_loan), ezek közül több értéket 0–1 bináris formátumra konvertáltam az egyszerűbb feldolgozhatóság érdekében. Például a transfer_type oszlopban az "Arrivals" és "Departures" értékeket numerikus formában jelöltem meg (1 és 0).

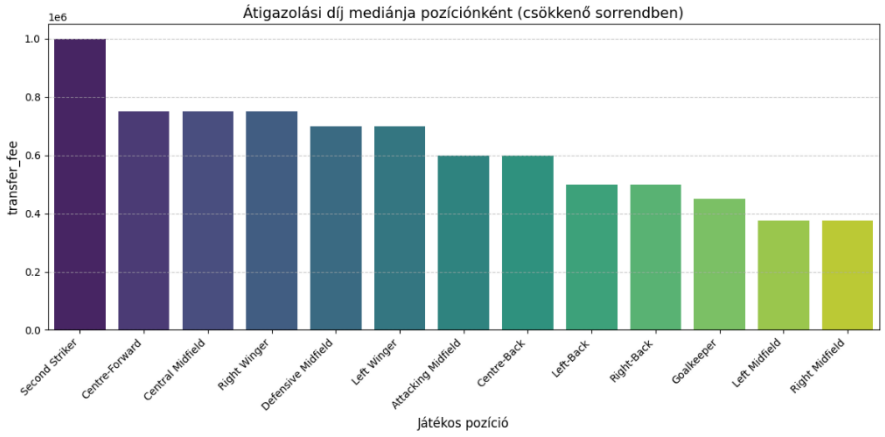
A player_position, league és más hasonló, szöveges kategóriákat hordozó oszlopokat is előfeldolgoztam. Valahol dummy változókra bontottam őket (one-hot encoding). Bizonyos esetekben viszont inkább ordinal encoding technikát alkalmaztam – például a player_age vagy a league_rank esetében –, ahol a kategóriáknak természetes sorrendisége is volt. [4]

Egyes oszlopokat – mint például az appearances, goals, minutes_played, in_squad – numerikus típusra konvertáltam, majd később ezeket külön is skáláztam (lásd 2.3). Mindezek mellett, mint ahogy már említettem, az oszlopok értékeinek eloszlását egyenként hisztogramokon keresztül vizsgáltam, hogy az adattisztítás megalapozottan történjen.

2.3 Feature engineering – új jellemzők létrehozása [1]

A modell hatékonyságának növelése érdekében több új jellemzőt is létrehoztam, amelyek segítettek jobban megragadni a játékosok piaci értékét befolyásoló tényezőket. Ezeket nem kizárólag technikai megfontolások alapján, hanem adatelemzési szempontból is megalapozott módon terveztem meg.

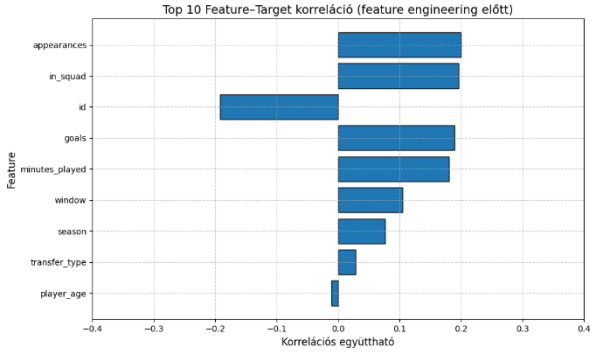
A 3. ábrán bemutatott medián átigazolási díjak pozíciónként például világosan megmutatták, hogy a különböző posztokon játszó játékosok értéke között szisztematikus eltérések vannak. Ez alapján alakítottam ki a position_score jellemzőt, amely a csatároknak magasabb, a védőknek alacsonyabb pontszámot rendelt.



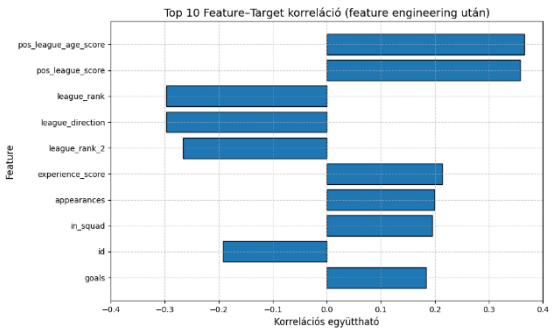
3. ábra: A különböző pozíciókhoz tartozó medián átigazolási díjak (csökkenő sorrendben)

További új jellemzők:

- *league_rank*: a célváltozó alapján rangsoroltam a bajnokságokat, így a modell képes lett figyelembe venni a liga presztízsét.
- *league_direction*: bináris mutatóként jeleztem, hogy a játékos erősebb vagy gyengébb bajnokságba igazolt.
- *age_score*: a játékosok életkorát ötszintű kategóriába soroltam (pl. very_young, prime, old).
- *experience_score*: az appearances és minutes_played alapján arányosítottam a játékos tapasztalatát.
- *pos_league_score* és *pos_league_age_score*: ezek az interakciós jellemzők a pozíció, liga szintje és életkor együttes hatását próbálták megragadni.



4. ábra: A célváltozóval legjobban korreláló jellemzők a feature engineering előtt



5. ábra: A célváltozóval legjobban korreláló jellemzők a feature engineering után

Egyes numerikus változókat külön skáláztam is, bár ez nem eredményezett szignifikáns teljesítményjavulást.

A fenti változók bevezetésének hatását korrelációs elemzéssel is vizsgáltam. A transfer_fee-hez való korrelációt figyelembe véve megállapítható, hogy a legtöbb új jellemző (pl. pos_league_age_score, league_rank, experience_score) erősebb kapcsolatot mutatott a célváltozóval, mint az eredeti, nyers változók. Ez jól szemléltethető a 4. és 5. ábrán, ahol a feature engineering előtti és utáni állapot közötti különbség látható: az újonnan képzett változók közül több is megelőzi az eredetieket a korrelációs rangsorban.
Ez alátámasztja, hogy a feature engineering valóban hozzájárult a modell prediktív teljesítményének javításához. Természetesn az egymásból képzett változók közül mindig csak egyet használtam fel a modellek tanítása során.

2.4 Célváltozó transzformálása

Az átigazolási díjak erősen jobbra ferde eloszlása miatt logaritmikus transzformációt [4] alkalmaztam a célváltozóra, azaz:

$$y_{log} = log(1 + transfer\_fee)$$

Ez a transzformáció csökkentette a kiugró értékek hatását, és javította a modellek tanulási képességét. A 2. ábrán látható hisztogram alapján, megfigyelhetjük, hogy valóban egyenletesebb eloszlást kaptunk.

2.5 Modellépítés és kiértékelés

A modellezés során három különböző regressziós modellt alkalmaztam [5]:

- Lineáris regresszió (LinearRegression): egy egyszerű, jól értelmezhető modell, amely feltételezi, hogy a célváltozó és a bemeneti jellemzők között lineáris kapcsolat áll fenn. Baseline modellként szolgált, amelyhez a komplexebb modellek teljesítményét viszonyítottam.
- Random Forest regresszor (RandomForestRegressor): egy ensemble alapú döntési fa algoritmus, amely több döntési fát tanít és azok átlaga alapján határozza meg az előrejelzést. Robusztus a túlillesztéssel szemben, és jól kezeli a nemlineáris kapcsolatokat, valamint a jellemzők közötti interakciókat.
- Gradient Boosting regresszor (GradientBoostingRegressor): szintén döntési fákra épülő ensemble modell, amely iteratív módon tanulja meg a hiba csökkentését. Gyakran kiváló prediktív teljesítményt nyújt, különösen komplex, strukturálatlan adathalmazokon.

A modelleket train_test_split [2] segítségével validáltam (80–20 arányban), és log-skálán tanítottam őket. Az előrejelzett értékeket expm1-tel visszaskáláztam, így a kiértékelés már az eredeti skálán történt.

A modellek kiértékelésére három jól ismert regressziós hibamértéket használtam:

- MAE (Mean Absolute Error): az előrejelzések átlagos abszolút eltérése a tényleges értékektől. Könnyen értelmezhető, robusztusabb a szélső értékekkel szemben.
- MSE (Mean Squared Error): a négyzetes eltérések átlaga. Érzékenyebb a nagyobb hibákra, így súlyozza a kiugró hibákat.
- MAPE (Mean Absolute Percentage Error): az abszolút hibák átlagos százalékos aránya. Jól mutatja, milyen arányban tévedett a modell a valós értékhez képest, de 0 közeli célértékeknél instabil lehet.

	Lineáris regresszió	Random Forest regresszor	Gradient Boosting regresszor
MAE	~1 800 000	~1 500 000	~1 800 000
MSE	~26 000 000 000 000	~21 000 000 000 000	~27 000 000 000 000
MAPE (%)	485.93	292.36	488.15

1. Táblázat: Modellek és kiértékelésük első körös eredményei

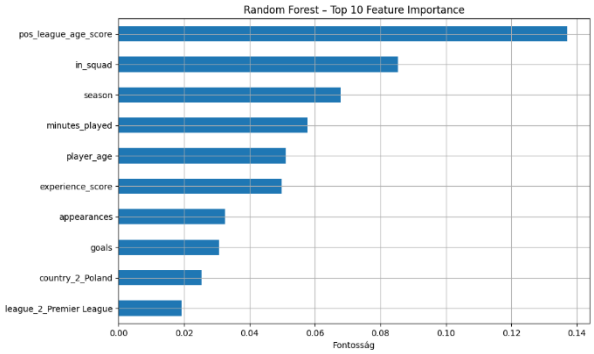
A MAE értékek (1,5–1,8 millió EUR) azt mutatják, hogy a modellek átlagosan ekkora összeget tévedtek az átigazolási díjak becslésénél. Ez azt jelenti, hogy például egy 8 millió eurós játékos árát a modell sok esetben 6,2–9,5 millió közé jósolta.

A MSE nagyságrendje miatt nehezebben értelmezhető (10^{13}–10^{14} körüli számok), mivel a hibákat négyzetes formában méri, így az extrém esetek torzíthatják.

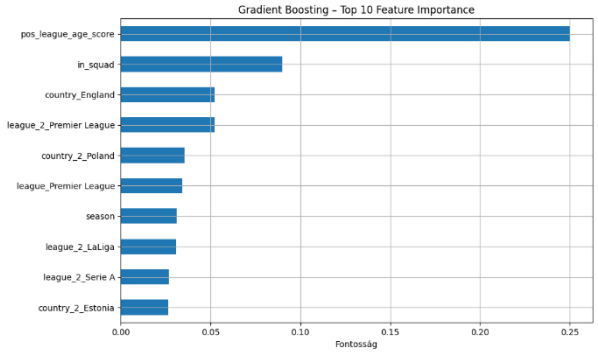
A MAPE (%) azt fejezi ki, hogy a modell átlagosan hány százalékkal tért el a valós értékektől. Itt például a Random Forest modell kb. 292%-os átlagos eltérést mutatott, míg a másik két modell még nagyobb százalékos hibát. Ennek oka, hogy sok játékos esetén az átigazolási díj viszonylag alacsony, így bármilyen tévedés nagy százalékos eltérést okoz.

Azt jól láthatjuk, hogy a Random Forest regresszor teljesített a legjobban, hiszen mindhárom mérőszámban ő produkálta a legjobb számakot.

A fenti modellek közül a Random Forest és a Gradient Boosting regresszor is lehetőséget biztosít arra, hogy az egyes bemeneti jellemzők fontosságát számszerűsítsük. A modell tanulási folyamata alapján meghatározható, hogy a döntésekhez mely attribútumok járultak hozzá a legnagyobb mértékben.



6. ábra: A Random Forest regresszor legfontosabb 10 jellemzője



7. ábra: A Gradient Boosting regresszor legfontosabb 10 jellemzője

A két modell jellemző fontossági értékei alapján jól látható, hogy a feature engineering során bevezetett pos_league_age_score változó messze a legnagyobb hatással bírt a becsült átigazolási díjakra. Továbbá az in_squad, season és minutes_played mezők is konzisztensen előkelő helyen szerepelnek, ami megerősíti ezek relevanciáját a célváltozó szempontjából

2.6 Kiugró értékek kezelése – felső és alsó 5% eltávolítása

Az előzetes modellezések során azt tapasztaltam, hogy az extrém magas vagy alacsony átigazolási díjak jelentős torzítást okozhatnak a tanítás során, különösen akkor, ha az adatok eloszlása nem szimmetrikus.

Ezért további adattisztítási lépésként az adathalmazt tovább szűkítettem oly módon, hogy a transfer_fee eloszlásának alsó és felső 5 százalékát eltávolítottam. A döntést a célváltozó eloszlásának vizuális vizsgálata és a metrikákban bekövetkező jelentős javulás is megerősítette.

Ezzel a lépéssel egy kiegyensúlyozottabb, az általános trendeket jobban reprezentáló tanítóhalmaz jött létre, amelyben kevesebb extrém eset dominálja a modellek tanulását.

	Lineáris regresszió	Random Forest regresszor	Gradient Boosting regresszor
MAE	~1 000 000	~900 000	~1 100 000
MSE	~4 000 000 000 000	~3 000 000 000 000	~4 000 000 000 000
MAPE (%)	138.87	99.46	138.00

2. Táblázat: Modellek és kiértékelésük kiugró értékek nélküli eredményei

2.7 Hiperparaméter-optimalizálás

Miután a három modell közül a Random Forest regresszor nyújtotta a legjobb teljesítményt a log-skálán tanított adatokon, további finomhangolást végeztem rajta hiperparaméter-keresés segítségével. [5] A cél az volt, hogy a modell általánosító képességét javítsam anélkül, hogy túlillesztés (overfitting)[1] lépne fel.

A hangoláshoz a GridSearchCV [2] eszközt használtam, amely lehetővé teszi, hogy több paraméterkombinációt kipróbáljak keresztvalidáció (cross-validation)[2] mellett.

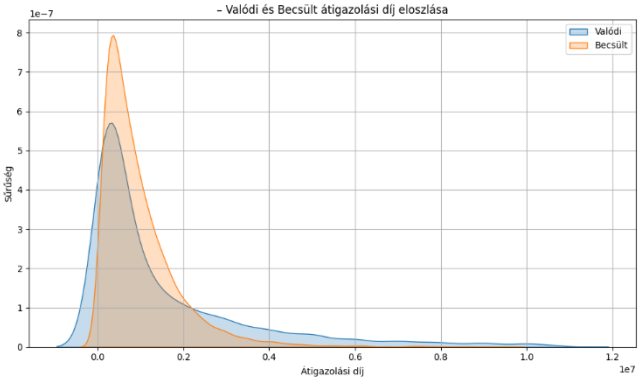
A keresési rácsot úgy állítottam össze, hogy a lehetséges kombinációk száma legfeljebb 24 legyen, így biztosítva, hogy a tanítási idő ésszerű keretek között maradjon.

A célfüggvényként a negatív MAE (Mean Absolute Error) értéket használtam, mivel ez közvetlenül tükrözi a modell predikciós pontosságát az átigazolási díjak skáláján.

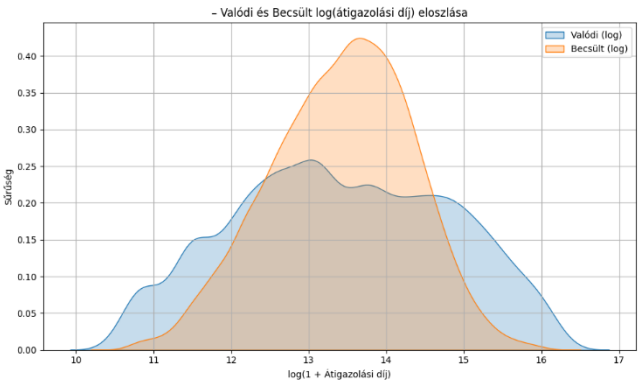
A keresés eredményeképpen sikerült enyhén csökkenteni a MAE értékét, miközben a MAPE is javult, ami arra utal, hogy a finomhangolt modell jobban tudott alkalmazkodni az adatok jellemzőihez.

	Random Forest regresszor
MAE	~850 000
MSE	~2 800 000 000 000
MAPE (%	90.33

3. Táblázat: Random Forest hiperparaméter-optimalizálás után



8. ábra: Valós és becsült átigazolási díjak eloszlása (natív skálán) Random Forest modell alapján



9. ábra: Valós és becsült átigazolási díjak eloszlása (log-skálán) Random Forest modell alapján

[1] Egy modell túlzottan alkalmazkodik a tanítóadatokhoz, így nem tud jól általánosítani új, ismeretlen adatokra.

[2] A keresztvalidáció egy olyan technika, amely során az adathalmazt több részre osztva többször újratanítjuk és teszteljük a modellt, hogy megbízhatóbban becsüljük meg annak általánosító képességét.

2.8 Összefoglalás

A félév során egy olyan gépi tanulási modell kialakításán dolgoztam, amely képes megbecsülni labdarúgó-játékosok átigazolási díját különböző, a játékosokra és az átigazolási helyzetre vonatkozó jellemzők alapján. A fő kérdés az volt, hogy az elérhető adatokból milyen mértékben lehet objektív becslést adni egy egyébként nagyrészt szubjektív, piaci tényezőktől is erősen befolyásolt értékre.

A cél egy prediktív modell létrehozása volt, amely a játékosok átigazolási díját becsli a pályán nyújtott teljesítményük, életkoruk, pozíciójuk és a klubjukhoz köthető ligák színvonala alapján. Az átigazolási piac egy rendkívül dinamikus és értékvezérelt környezet, ezért érdekes kérdés, hogy adatvezérelt módszerekkel mennyire lehet ezt a „piaci értéket” előre jelezni. Ez nemcsak sportgazdasági, hanem adattudományi szempontból is izgalmas kihívás.

A megközelítésem a klasszikus data science pipeline-t követte: az adatok betöltését és tisztítását követően különféle jellemzőket (feature-öket) konstruáltam, bináris és ordinális kódolásokat alkalmaztam, a célváltozót logaritmikus skálára transzformáltam a kiugró értékek kezelése érdekében, majd több regressziós modellt (lineáris regresszió, random forest, gradient boosting) tanítottam és értékeltem. Hiperparaméter-hangolást is végeztem a legjobban teljesítő modellre.

A modell legjobb teljesítményét a Random Forest regresszor adta, különösen akkor, amikor új, összetett jellemzőket (pl. pos_league_age_score) is bevontam. A kiugró értékek eltávolítása, a log-skálás feldolgozás, valamint a megfelelő jellemzők kiválasztása jelentősen javította a becslések pontosságát. A MAE érték ~850 ezer euró körül alakult, a MAPE pedig 90% volt a legjobb modell esetében, ami egy ilyen szélsőséges és heterogén adathalmaznál elfogadható eredménynek tekinthető.

A projekt rávilágított arra, hogy még egy ilyen komplex és sok bizonytalansággal terhelt probléma esetén is lehet értelmezhető predikciókat adni adatvezérelt módszerekkel. Emellett megerősített abban, hogy a feature engineering kulcsfontosságú lépés, amely sok esetben többet számít, mint a választott modell típusa. A munka során gyakorlati tapasztalatot szereztem a teljes adatfeldolgozási és modellezési folyamatban, és számos új szempontot ismertem meg, amelyeket a jövőbeli hasonló projektekben is hasznosítani tudok majd.

A modell továbbfejlesztése szempontjából több irány is nyitott. Egyrészt érdemes lehetne más fajta adatokat is felhasználni különböző adatforrásokból. Továbbá a jelenlegi modell featureinek kiválasztása során használhatnánk Forward Selection technikát, ami még pár százalékot javíthatna a modellünkön. Emellett, ha több év adatát tekintjük át időbeli sorrendben, akkor akár idősoros előrejelzésekkel is lehetne kísérletezni.

3. Irodalom, és csatlakozó dokumentumok jegyzéke

A tanulmányozott irodalom jegyzéke:

[1] Aurélien Géron: *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, O'Reilly Media, 2nd Edition, 2019. ISBN: 9781492032649.

[2] *scikit-learn Machine Learning in Python*, https://scikit-learn.org/stable/index.html, Utolsó olvasá időpontja: 2025.05.07.

[3] *B.Sc, VITMAL01 Informatika önállólabor a BME TMIT-n, heti 4 óra.* , https://inflab.tmit.bme.hu/25t/bsc.shtml, szerk.: Németh Felicián, 2025. május 5.

[4] *07_ensemble_learning_and_random_forests.ipynb,* https://colab.research.google.com/github/ageron/handson-ml3/blob/main/07_ensemble_learning_and_random_forests.ipynb#scrollTo=sR73MrmQJw7y, Utolsó olvasás időpontja: 2025.05.07.

[5] *02_end_to_end_machine_learning_project.ipynb,* https://colab.research.google.com/github/ageron/handson-ml3/blob/main/02_end_to_end_machine_learning_project.ipynb#scrollTo=QO8noQEGcl09, Utolsó olvasás időpontja: 2025.05.07