OZNAL - Dátová analýza dát hráčov z hry FIFA 19

Timotej Zaťko and Tomáš Hoffer

Fakulta informatiky a informačných technológií STU v Bratislave Ilkovičova 2, 842 16 Bratislava 4
xzatkot1@stuba.sk xhoffer@stuba.sk
https://www.fiit.stuba.sk/

Abstract. Táto práca obsahuje analýzu dát hráčov z hry FIFA 19, opis dát a ich charakteristiky. V tejto práci skúmame vzťahy atribútov na predikovaný atribút trhovej ceny hráča a jeho hernú pozíciu na ihrisku.

Keywords: Analýza dát, FIFA 19, Strojové Učenie, Regresia, Klasifikácia

1 Opis problému a motivácia

Rozhodli sme sa pre analýzu dát hráčov z hry FIFA 19. V hre sa nachádza veľké množstvo hráčov z rôznych krajín, hrajúcich v rôznych súťažiach. Ich schopnosti v hre by mali odzrkadlovať ich schopnosti z reálneho sveta. Tvorcovia hry sa o to snažia vytvorením herných atribútov, akými sú napríklad rýchlosť šprintu, sila strely, zakončovanie alebo hlavičkovanie ktoré dokázali vyjadriť číselne. V projekte sa budeme snažiť na základe týchto atribútov klasifikovať hráčov do hernej pozície a predikovať ich trhovú hodnotu v hre. Keďže hra hráčom neukazuje vždy trhovú hodnotu hráča, ale len jeho atribúty, náš model môže byť užitočný pri určovaní výšky ponúkanej sumy za prestup hráča pri vyjednávaní v hre. Keďže v reálnom svete schopnosti hráčov (napr. zakončovanie) niesú nijako numericky vyjadrené skúsime využiť to, že v hre vyjadrené sú, na zistenie, ktoré atribúty sú dôležité pre určité herné pozície.

2 Opis dát s charakteristika dát

Naša dátová sada obsahuje 18207 záznamov - hráčov, ktorý každý z nich má 87 atribútov. Z toho je 42 numerických a 45 kategorických.

2.1 Očistenie dát

Kvôli analýze bolo treba dátovú sadu, očistiť a urobiť predspracovanie niektorých atribútov. Bolo nutné spraviť nasledovné úpravy – konverzia a očistenie finančných hodnôt (napr. ‘$77.5M‘, ‘$1K‘), konverzia mier (napr. ‘159lbs‘, ‘5’11‘),

konverzia časových údajov (napr. ‘Jan 25, 2019‘, ‘2018‘) a rozdelenie niektorých atribútov na viac atribútov - niektoré atribúty obsahovali dva numerické atribúty. Taktiež atribút určujúci pozíciu hráča obsahoval až 38 rôznych herných pozícii, preto sme sa rozhodli tento atribút rozšíriť do ďalších dvoch atribútov v ktorých sme podobné pozície spojili čím vznikli dva atribúty s 13 resp. 4 hodnotami. Po týchto úpravách sa nám počet atribútov zmenil na 129 – 89 numerických a 40 kategorických atribútov.

Toto veľké množstvo atribútov sme zaradili do nasledovných kategórií, uvádzame k nim aj niektoré atribúty. Atribúty sme rozdelili aj na základe toho či jeho hodnotu musela hra nejakým spôsobom odvodiť z reálneho sveta.

- **človek** - meno, národnosť, výška, hmotnosť, vek...
- **futbalový hráč (reálny svet)** - názov klubu, číslo dresu, dátum príchodu do klubu, dĺžka kontraktu, plat hráča, hosťovský klub, herná pozícia, preferovaná noha
- **futbalový hráč (hra)** - trhová hodnota hráča, triky, pracovitosť v útoku/obrane, vhodnosť na určitú špecifickú pozíciu a potenciálny rast (78 atribútov), medzinárodná reputácia
- **futbalové schopnosti hráča určené hrou** (celkovo 34 atribútov) - krátke prihrávky, hlavičkovanie, zakončovanie...
- **iné atribúty hry** - logo klubu, vlajka (na základe národnosti), typ postavy (tj. typ herného modelu), typ tváre (tj. kvôli herného modelu)

2.2 Analýza chýbajúcich hodnôt

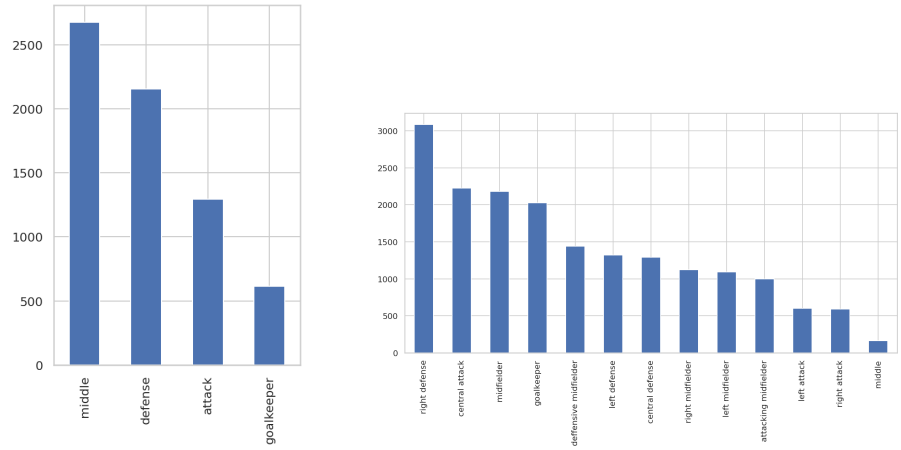
Ako prvé sme sa pozreli na chýbajúce hodnoty v našej dátovej sade. Početnosti sme vizualizovali stĺpcovým diagramom a vzťahy tepelnou mapou a dendrogramom. Zistili sme, že niektorým hráčom chýbajú hodnoty v atribútoch ako sú klub, plat, dĺžka kontraktu, číslo dresu či výkupná klauzula. Tieto hodnoty chýbajú väčšinou spoločne, o čom sme sa presvedčili v dendrograme a tepelnej mape. Je to aj logické, keďže hráč, ktorý nemá klub nemôže poberať plat alebo mať výkupnú klauzulu v kontrakte. Z našich zistení môžeme konštatovať, že ak hráčovi chýbajú hodnoty v atribúte, väčšinou sa nejedná o chybu v úplnosti dátovej sady.

2.3 Analýza z pohľadu pozície hráča

Analýzu sme ďalej realizovali z pohľadu dvoch atribútov, ktoré sa budeme snažiť predikovať - pozíciu hráča (kategorický atribút) a jeho trhovú hodnotu (numerický atribút). Dátová sada celkovo obsahuje 36 herných pozícií.

Na základe našich futbalových znalostí sme sa rozhodli zoskupiť podobné pozície do skupín, čím sme znížili počet rôznych hodnôt v atribúte ‘Position‘ (Fig. 1). Môžeme pozorovať, že triedy nie sú vyvážené.

Pozíciu sa môžeme pokúsiť predikovať na základe 34 atribútov definujúcich futbalové schopnosti hráča určené hrou. Vizualizovali sme všetky dvojice týchto atribútov a pomocou ”scatter plot” -ov sme hľadali zhluky jednotlivých pozícií.



(a) 4 herné pozície(‘Position (4)‘)

(b) 13 herných pozícií (‘Position (13)‘)

Fig. 1: Početnosti nových atribútov ‘Position (4)‘ a ‘Position (13)‘ po zoskupení podobných hodnôt z atribútu ‘Position‘.

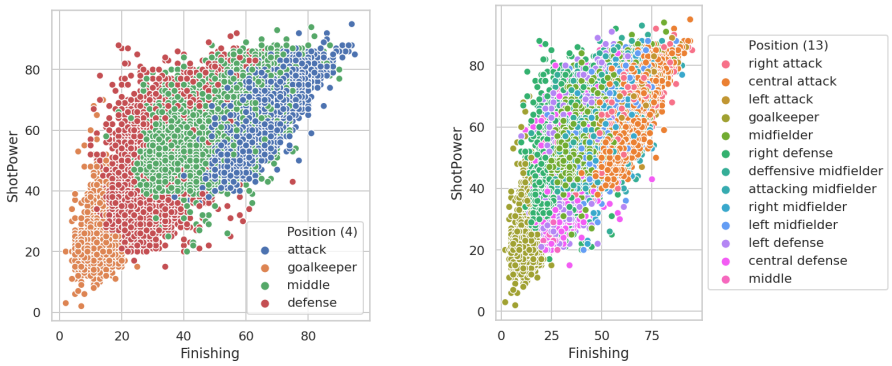
Ako príklad uvádzame vzťah medzi atribútmi ‘Shot Power‘ (sila strely) a ‘Finishing‘ (zakončovanie), kde môžeme pozorovať jednotlivé zhluky podľa pozície hráča (Fig. 2).

Z prieskumnej analýzy tiež vyplýva, že pre konkrétne pozície hráčov sú typické určité čísla dresov. Pre brankárov (angl. goalkeeper) je typickým číslom dresu číslo *1* pričom toto číslo nemá priredený žiaden hráč na inej pozícii. Pre útočníkov (angl. attack) to je *9*, pre obrancov (angl. defense) sú to čísla *2* - *6* a pre stredopoloarov *7*, *8* a *10*. Tento atribút môže byť veľmi dobrý na klasifikáciu pozície hráča, avšak my sa v prvom rade zamieriame na predikciu z herných atribútov. Atribúty, ako napríklad čislo dresu nám môžu úlohu príliš zjednodušiť.

Ďalšou zaujímavosťou je, že v našej dátovej sade sa nachádzajú prevažne hráči ktorí preferujú pravú nohu, avšak na pozícii ľavého obrancu výrazne prevládajú hráči s preferovanou ľavou nohou. (Fig. 3).

2.4 Analýza z pohľadu trhovej hodnoty hráča

Pomocou Pearsonovho korelačného koeficientu sme hľadali korelácie medzi atribútom ‘Value‘ a ostatnými numerickými atribútmi. Takmer lineárnu koreláciu (0.99) vykazuje atribút ‘Release Cause‘ (Fig. 4). Hráči s vysokou trhovou hodnotou majú v dátovej sade podpísanú zmluvu s vyššou výkupnou klauzulou. Vysokú koreláciu taktiež vykazujú atribúty ‘Overall‘ (0.631), ‘Wage‘ (0.850) a ‘International Reputation‘ (0.656).



(a) 4 herné pozície (‘Position (4)‘) (b) 13 henrých pozícií (‘Position (13)‘)

Fig. 2: Vzťah medzi atribútmi ‘Shot Power‘ a ‘Finishing‘. Pre 4 herné pozície sú zhluky zreteľnejšie ako pre 13 herných pozícií.

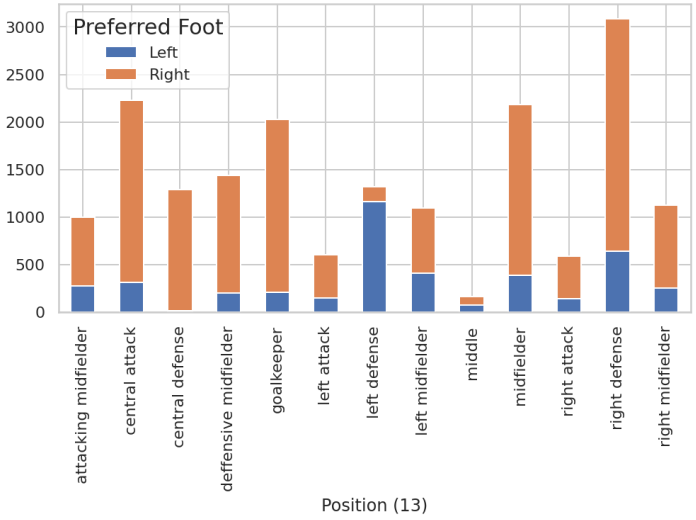
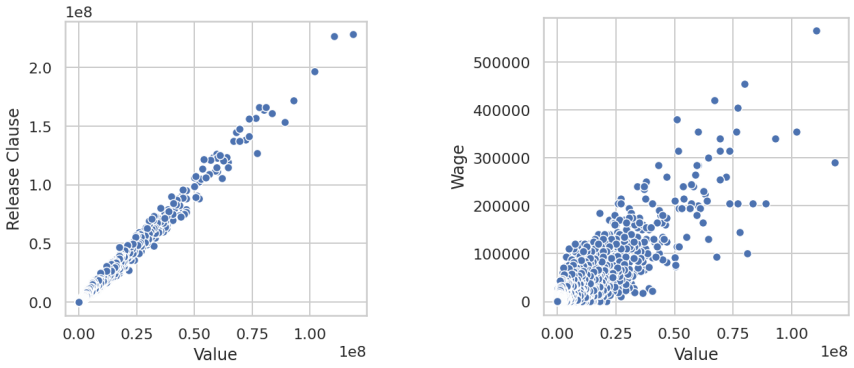


Fig. 3: Preferovaná noha hráčov podľa hernej pozície.



(a) Vzťah ‘Release Cause‘ voči ‘Value‘.

(b) Vzťah ‘Wage‘ voči ‘Value‘.

Fig. 4: Niektoré atribúty vykazujú vysokú mieru korelácie s atribútom ‘Value‘.

3 Definovanie úlohy objavovania znalostí

Rozhodli sme sa, že budeme vykonávať nasledujúce úlohy:

- predikcia hernej pozície hráča – všeobecnej (4 triedy), rozšírenej (13 tried) – *klasifikácia*
- predikcia hodnoty hráča (atribút s názvom ‘Value‘) – *regresia*

Obe tieto úlohy budeme realizovať z atribútov určujúcich herné shopnosti hráča a nie zo všetkých atribútov, keďže niektoré atribúty výrazne ovplyvňujú predikovanú premennú a to – Release Clause - Value; Jersey Number - Position.

4 Predpokladaný scenár riešenia (problémy)

Predpokladáme, že bude potrebné vykonať nasledovné úlohy:

- predspracovanie kategorických hodnôt (tj. one-hot encoding)
- normalizácia dát
- over-sampling
- trénovanie modelu

Trénovanie modelu zahŕňa výber atribútov (angl. feature selection) a výber a trénovanie modelu. Na úlohu predikcie hodnoty hráča budeme pravdepodobne používať lineárnu regresiu/neurónovú sieť a na určenie hernej pozície hráča rozhodovací strom / náhodný les / SVM / neurónovú sieť.

5 Predspracovanie dát

Prvotnému predspracovaniu dát na účely analýzy sme sa venovali v kapitole 2.1 kde sme vykonali nasledovné transformácie dát:

- Transformácia hodnôt peňažných atribútov (napr. hodnoty '1.2 M', '200 K' sme transformovali na jednotný numerický tvar)
- Dátumové atribúty sme transformovali do jednotného formátu, reprezentovaného UNIX-ovou časovou pečiatkou
- Atribúty miery (napr. výška a hmotnosť hráča) sme previedli do rovnakých jednotiek.
- Špeciálne pozície hráča boli udávané v tvare reťazca *Value+Grow*. Atribúty špeciálnych pozícií sme rozdelili na dvojice aby nám vznikli numerické atribúty.
- Atribút *Work Rate* bol udávaný reťazcom v tvare *Attack/Defense*. Tento atribút sme tiež rozdelili na 2 numerické.
- Boolean atribúty udávané reťazcami sme transformovali na numerické 0/1
- Vytvorili sme 2 nové atribúty Position(4) a Position(13), ktoré reprezentujú pozíciu hráča po zoskúpení pozícií, opísanom v kapitole 2.3
- Vytvorili sme atribút *Contract length*, ktorý reprezentuje dĺžku aktuálne podpísanej zmluvy hráča

Kategorické atribúty sme transformovali metódou "One-Hot Encoding". Numerické atribúty sa nachádzali v rôznych rozsahoch a jednotkách (napríklad hmotnosť v lb, výška v cm, rôzne schopnosti hráča v rozsahu $< 0, 100 >$). Rozhodli sme sa ich normalizovať do jednoetného intervalu $< 0, 1 >$.

5.1 Chýbajúce hodnoty

V kapitole 2.2 sme opísali dôvod výskytu chýbajúcich hodnôt. Chýbajúce hodnoty opísané v kapitole 2.2 považujeme za opodstatnené a nebudeme ich nahrádzať. Tj. dané hdonoty chýbajú pretože majú chýbať a nie preto, že by boli dáta neúplné.

5.2 Vychýlené hodnoty

Rozdelenie číselných hodnôt v dátovej sade sme podrobne analyzovali a vizualizovali. Vychýlené hodnoty sú iba v niektorých atribútoch.

5.3 Problém nevyváženosti tried

Pri zoskúpení všetkých pozícií do 13 tried sme odhalili nevyváženosť jednotlivých tried (Fig.1), čo by mohlo negatívne ovplyvniť presnosť predikcie. Tento problém sme sa rozhodli riešiť metódou nazývanou *oversampling* a konkrétne algoritmom SMOTE [1], ktorý vytvára syntetické inštancie minoritných tried/y pomocou lineárnej kombinácie reálnych inštancií.

6 Výber atribútov

Predikciu pri oboch typoch úloh budeme realizovať pomocou 34 atribútov definujúcich futbalové schopnosti hráča určené hrou. Využijeme tiež demografické údaje, napr. výšku, hmotnosť a vek. K tejto základnej množine atribútov sme sa dopracovali na základe doménovej znalosti a rozsiahlej prieskumnej analýzy. V častiach 2.3 a 2.4 sa venujeme výberu týchto atribútov a opisujeme dôvody, prečo sme niektoré atribúty do tejto množiny nezahrnuli. Pri riešení klasifikačnej úlohy rozšírime základnú sadu atribútov o atribúty *Preferred Foot*, *Work Rate Attack* a *Work Rate Defense*. Pri regresnej úlohe využijeme dopočítaný atribút *Contract Length* a využijeme tiež medzinárodnú reputáciu hráča (*angl. International reputation*). Množinu vybraných atribútov sa pravdepodobne ešte pokúsime zredukovať (angl. feature selection). V závislosti od modelu zvolíme konkrétny prístup. Niektoré modely podporujú tzv. *feature importances*. Pri modeloch, ktoré to nepodorujú môžeme použiť rekurzívnu elimináciu atribútov alebo prírodou inšpirované algoritmy, ktoré v tejto oblasti dosahujú veľmi dobré výsledky [6].

7 Príbuzná práca

V našej doméne (avšak na dátach zo staršej verzie hry) Soto-Valero, C. [4] zisťoval, ktoré atribúty sú dôležité pre jednotlivé herné pozície. Využil PCA na zredukovanie dimenzionality zo všetkých herných atribútov a následne zhlukovaním klasifikoval hráčov do jednotlivých herných pozící. Následne pomocou algoritmu *Gradient tree boosting* ohodnotil dôležitosť jendotlivých atribútov. Na klasifikovanie využil takzvané učenie bez učiteľa, čím sa jeho práca líši od tej našej.

Nazim R. a spol. sa snažia predikovať pozíciu hráča na základe podobných atribútov, avšak využívajú dátovú sadu zo športovej školy *Bukit Jalil Sports School*. Na klasifikáciu využívajú bayesovské siete, rozhodovacie stromy a algoritmus najbližšieho suseda. Hráči boli klasifikovaní do 10 tried. Vzhľadom na to, že dátova sáda obsahovala len 100 prvkov sa autori rozhodli použiť *"leave-one-out"* validáciu [3].

Yaldo L. a Shamir L. sa snažia riešiť problém predikcie týždenného platu futbalového hráča pomocou 8 rôznych algoritmov strojového učenia na podobnej dátovej sade FIFA z roku 2016 [5]. Autori na predikciu využili napríklad aditívnu regresiu či rozhodovacie stromy. Platy hráčov sú predikované na základe ich pozície, ale aj atribútov hry, ktoré charakterizujú ich schopnosti. Zaujímavým zistením je fakt, že platy hráčov nezávisia len od ich schopností a pozície hráča, ale aj celkovej reputácie a jeho popularity. Známy hráč L. Messi by napríklad pri predikcii len na základe jeho schopností nedosiahol ani polovičný plat. Rozhodli sme sa teda medzinárodnú reputáciu hráča zahrnúť medzi črty použité na predikciu. Autori používajú Piersonov korelačný koeficient (CC) a priemernú absolútnu odchýlku (MAE) na meranie úspešnosti.

8 Výber metrík pre evaluáciu úspešnosti

Klasifikácia – keďže klasifikujeme do 4 resp. 13 tried, ktoré sú nevyvážené, budeme sledovať metriky **f1 micro** a **f1 macro**. Táto metrika je harmonickým priemerom *precision* a *recall* [2]. Nás zaujíma predovšetkým varianta **micro**, keďže máme nevyvážené triedy a chceme aby metrika brala do úvahy početnosti tried (tj. náš model nemusí byť taký dobrý v klasifikácii minoritných tried). Budeme tiež sledovať aj variantu **macro**, ktorá berie do úvahy všetky triedy rovnako.

Regresia – pri regresii budeme sledovať metriku s názvom **RMSE** (Root Mean Sqaure Error).

9 Opis použitých algoritmov

9.1 Predikcia pozície hráča (klasifikácia)

Rozhodovací strom – klasifikátor založený na stromovej štruktúre. Uzly reprezentujú test vykonaný nad atribútom inštancie a vetvy reprezentujú možné výsledky testu. Rozhodovací strom pri klasifikácii začína v koreni a postupne prechádza cez jednotlivé ulzy až k listu, ktorý reprezentuje výsledok klasifikácie (triedu).

9.2 Predikcia trhovej hodnoty hráča (regresia)

Lineárna regresia – algoritmus, ktorý sa pokúša modelovať vzťah medzi dvoma premennými pomocou lineárnej funkcie. Rovnica lineárnej regresie má tvar $Y = kX + q$, kde Y a X sú závislé premenné, q udáva posun a k udáva sklon funkcie. Algoritmus je vhodný pri riešení regresných úloh, pri ktorých modelujeme lineárny vzťah medzi premennými. Vzájomnú koreláciu je vždy vhodné vopred overiť pomocou Pearsonovho korelačného koeficientu.

9.3 Skupinové učenie (angl. "ensemble learning")

Náhodný les – klasifikátor/regresor, ktorý využíva viacero rozhodovacích stromov, ktoré sú trénované nezávisle a každý pri klasifikácii dostane 1 hlas. Náhodný les následne vyberie triedu s najvyšším počtom hlasov. Výhodou algoritmu je tiež zabránenie korelácie medzi jednotlivými rozhodovacími stromami vďaka využitiu "bagging" metódy. Rozhodovacie stromy sú trénované na množinách dát, ktoré vznikajú výberom s nahradzovaním (angl. "selection with replacement"). Vďaka tejto metóde výberu vznikajú rozhodovacie stromy rôznych šírok a hĺbok a práve ich rôznorodosť prispieva k úspešnosti algoritmu náhodného lesa.

Adaboost – klasifikátor/regresor založený na lineárnej kombinácii viacerých tzv. slabých klasifikátorov/regresorov, trénovaných na iteratívne modifikovaných dátach ("boosting" iteráciách). Iterácie pozostávajú z aplikácie váh $w_1, w_2, ...w_N$, pričom N reprezentuje počet prvkov. Počas každej iterácie sú váhy nesprávne klasifikovaných prvkov z predchádzajúcej iterácie zvýšené, resp. znížené pri správnej klasifikácii. Pozorovania, ktoré je náročné správne klasifikovať postupne dostanú vysokú váhu a slabé klasifikátory sa pri ďalšom trénovaní zameriavajú na tieto pozorovania. Predikcie slabých klasifikátorov sú nakoniec kombinované a použité na finálnu predikciu.

Gradient tree boosting – klasifikátor/regresor využíva rozhodovacie stromy ako slabé klasifikátory/regresory. Rozhodovacie stromy sú postupne vytvárané a pridávané do množiny použitých slabých klasifikátorov s využitím metódy klesajúceho gradientu (angl. "gradient descent") na minimalizáciu chybovej funkcie.

Voting classifier/regressor – klasifikátor/regresor kombinujúci výsledky viacerých rôznych klasifikátorov/regresorov. Použitie algoritmu je vhodné napríklad v prípade, keď máme viacero úspešných modelov na vyváženie ich slabostí. Metóda "hard voting" priradzuje každému modelu hlas s rovnakou váhou. Metóda "soft voting" umožňuje priradenie váhy hlasom jednotlivých modelov.

Stacked generalization – klasifikátor/regresor kombinujúci výsledky viacerých modelov. Výsledky modelov sú vstupom finálneho modelu, ktorý je trénovaný s využitím krížovej validácie.

10 Experimenty

10.1 Predikcia pozície hráča (klasifikácia)

Úlohu sme sa v prvej etape rozhodli riešiť pomocou viacerých jednoduchých modelov. Klasifikácia hráča do 4 pozícií už pri základných nastaveniach modelov vykazovala vysokú úspešnosť 1. Model sa najviac mýlil pri klasifikácii do triedy *middle*. Je to pochopiteľné, keďže stredopoliari sú "medzi" útočníkmi a obrancami 5.

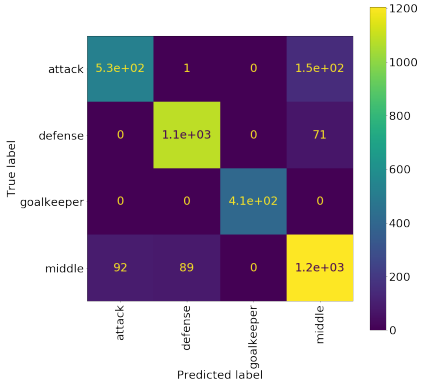
Pri klasifikácii do 13 tried pozície je úspešnosť jednoduchých modelov významne nižia a niektoré triedy jednotlivé modely dokonca vôbec nepredikovali.

10.2 Predikcia trhovej hodnoty hráča (regresia)

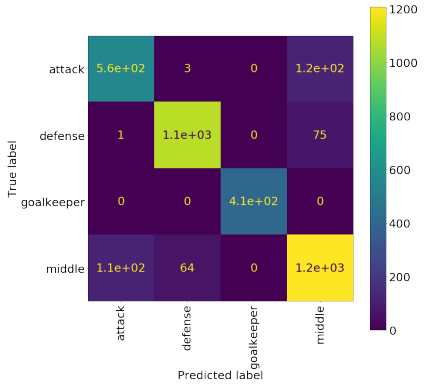
Na riešenie tejto úlohy sme využili lineárnu regresiu. V tejto úlohe sa nám zatiaľ nepodarilo dosiahnuť výraznejšie výsledky (Obr. č.6).

Model	Position (4) – f1 skóre						Position (13) – f1 skóre	
	Attack	Defense	GK	Middle	Micro	Macro	Micro	Macro
Rozhodovací strom	0.74	0.87	1	0.76	0.81	0.84	0.55	0.42
Logistická regresia	0.82	0.94	1	0.87	0.87	0.90	0.71	0.53
SVM	0.81	0.93	1	0.84	0.88	0.89	0.66	0.44
Náhodný les	0.81	0.93	1	0.86	0.89	0.89	0.68	0.50

Table 1: Úspešnosti jednoduchých modelov pri predikcii 'Position 4' a 'Position 13'. Pre predikciu štyroch herných pozícií zobrazujeme aj metriku **f1** pre jednotlivé pozície. V oboch prípadoch bola najlepšia logistická regresia a náhodný les.



(a) Matica zmätenia pre model náhodny les.



(b) Matica zmätenia pre model logistickej regresie.

Fig. 5: Matice zmätenia pre dva z našich natrénovaných modelov, môžeme si všimnúť že model robí chyby pri klasifikácii do triedy *middle*.

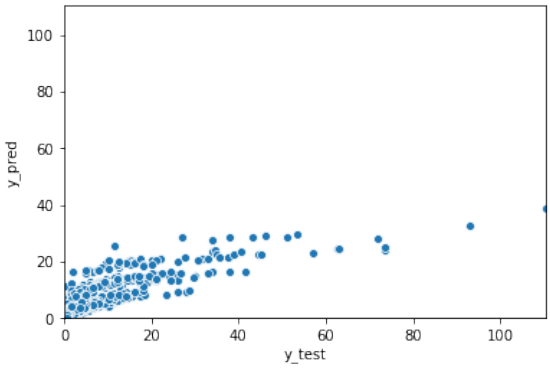


Fig. 6: Predikované hodnoty hráčov a reálne hodnoty hráčov v miliónoch. Zatiaľ sa nám nepodarilo dosiahnúť lepšie výsledky ($RMSE = 3{,}560331$).

10.3 Oversampling

Keďže pri našej klasifikačnej úlohe sú jednotlivé triedy nevyvážené (Sekcia 5.3) pokúsili sme sa zvýšiť početnosti minoritných tried. Vyskúšali sme minoritné triedy dovzorkovávať náhodne a zároveň metódou SMOTE a jej variáciami (SVM-SMOTE, BorderlineSMOTE). Tiež sme skúšali rôzne kombinácie hyperparametrov týchto metód. Ako model s pouižili náhodný les, keďže zo všetkých modelov sa trénoval najrýchlejšie a patril medzi najlepšie modely (Tab. 1). Pri klasifikácii do 'Position 4' a 'Position 13' oversampling modely výrazne nezlepšil. Pri všetkých modeloch sa zlepšila metrika *f1 macro* a zas shoršila *f1 micro*, tj. model sa zlepšil na klasifikácii minoritných tried na úkor mierneho zhoršenie klasifikácie tých majoritných. Najlepší model s použitím SMOTE bol lepší o 2 percentuálne body v metrike *f1 macro* a 0.4 percentualného bodu horší v metrike *f1 micro*. Použitím oversampling-u sme dosiahli iba mierne zlepšenie v klasifiácii.

10.4 Výber atribútov (angl. feature selection)

Jednotlivé atribúty sme zatiaľ vyberali iba na základe doménových znalostí a podľa nami definovanej úlohy objavovania znalostí (Sekcia 3). Keďže je týchto atribútov stále veľa, pokúsili sme sa ich zredukovať aby sme mali jednoduchší a ľahšie interpretovateľný model. Vyskúšali sme niekoľko spôsobov, a to konkrétne: odstránenie atribútov s nízkou varianciou, výber atribútov na základe jednorozmerných štatistických testov (angl. "univariate feature selection"), rekurzívnu elimináciu atribútov, výber na základe ich dôležitosti pre model (angl. "feature importances"). V rámci týchto metód sme skúšali rôzne parametre a výber atribútov sme vykonávali pre klasifikačnú úlohu (klasifikátor - náhodný les) a aj pre regresnú úlohu (lineárna regresia).

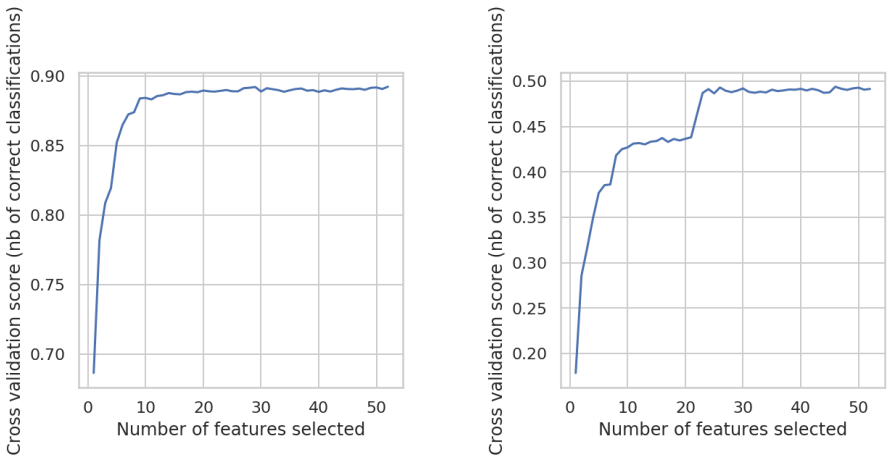
Výber atribútov výrazne nezlepšil úspešnosť modelu, ale ani nezhoršil. Použitím polovice zo všetkých atribútov sme dosiahli rovnako dobré výsledky ako keby sme ich použili všetky (Obr. 7). Pri regresii bol tento trend podobný.

Výberom atribútov sa nám podarilo natrénovať jednoduchšie modely s podobnou úspešnosťou ako bez výberu atribútov.

10.5 Ensemble

Vyskúšali sme niektoré ensemble metódy, pomocou ktorých sa nám podarilo dosiahnuť ešte lepšie výsledky ako sme dosiahli doteraz. Vyskúšali sme AdaBoost, Gradient Tree Boosting (ďalej GDT), Voting (hard a soft) kde sme použili logiskickú regresiu, rozhodovací strom a SVM. Ďalej sme vyskúšali metódu stacked generalization, kde sme použili náhodný les, AdaBoost a GDT. S metódami sme viac do hĺbky neexperimentovali. Pri prvotých experimentoch bol najlepší GDT s tým, že jeho trénovanie trvalo pomerne krátko, čo nám umožnilo použiť ho aj pri hyper-parameter tunningu.

Pri regresnej úlohe spomedzi použitých metód bol výrazne najlepší Gradient Tree Boosting s pomerne krátkym časom trénovania.



(a) Úspešnosť pri klasifikácii 'Position 4' podľa počtu použitých atribútov

(b) Úspešnosť pri klasifikácii 'Position 13' podľa počtu použitých atribútov

Fig. 7: Úspešnosti pri klasifikácii podľa počtu použitých atribútov. Graf vznikol použitím rekurzívnej eliminácie s krížovou validáciou (*k=2*) na klasifikátore náhodný les.

10.6 Hyper-parameter tunning

Pri oboch úlohách sme realizovali hyper-parameter tunning pomocou algoritmu Grid Search na metóde Gradient Tree Boosting. V oboch prípadoch sme optimalizovali nasledujúce hyper-parametre - *min samples leaf*, *number of estimators*, *max depth*, *min samples split* a *subsample* (tieto parametre sú bližšie vysvetlené tu). Dosiahnuté výsledky sme vizualizovali pomocou tepelných máp. Keďže je grid search najmenej optimálna metóda hľadania hyper-parametrov, prehľadávali sme pomerne malý priestor parametrov. Taktiež sme sa pokúšali hľadať najoptimálnejšie dvojice parametrov alebo sme optimalizovali iba jeden parameter. Aj preto výsledky nie sú výrazne lepšie.

Pri optimalizácii hyper-parametrov sme použili krížovú validáciu s $k = 5$ (tj. dátová sada bola rozdelená na 5 častí).

Takto sa nám pre obe úlohy podarilo zlepšiť natrénované modely, avšak niektoré len minimálne (Tab. 2).

S vybranými najoptimálnejšími parametrami sme následne experimentovali s oversamplingom, ktorý model vždy zhoršil. Taktiež sme sa pokúsili natrénovať čo najjednoduchší model - tj. použiť čo najmenej atribútov, taký model je potom ľahšie interpretovateľný. V prípade klasifikácie do 'Position 4' sa použitím iba 13 atribútov skóre zhoršilo iba minimálne z *0.896/0.906* (f1 micro/macro) na

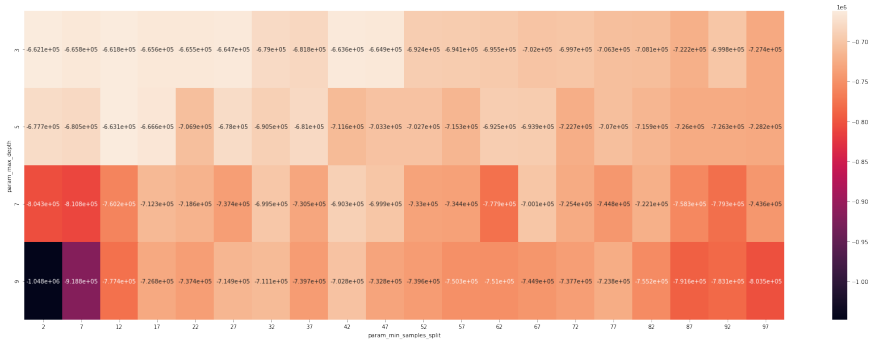


Fig. 8: Tepelná mapa skóre natrénovaného modelu (regresia) podľa zvolených parametrov. Na osi x je parameter *min samples split* na osi y je parameter *max depth*. Farba vyjadruje skóre modelu (neg. RMSE). Čím je farba bledšia, tým je chyba modelu nižšia.

Úloha	Predikovaný atribút	Model	Metrika	Pred	Po
Klasifikácia	Position (4)	GBT	F1 micro/macro	0.893/0.904	0.896/0.906
Klasifikácia	Position (13)	GBT	F1 micro/macro	0.698/0.526	0.696/0.535
Regresia	Value	GBT	RMSE	0.529 M	0.506 M

Table 2: Porovnanie úspešnosti modelov pred a po optimalizácii hyperparametrov.

0.873/0.887. V prípade klasifikácie do ‘Position 13‘ bol tento trend podobný. V prípade regresie sa nám dokonca podarilo natrénovať lepší model s *RMSE=0,474* (Obr. 9) z pôvodných *0.506* s použitím 33 zo 71 atribútov. S použitím iba dvoch atribútov - veku hráča a jeho celkového hodnotenia (stĺpec ‘Overall‘) sme dosiahli *RMSE=1*, teda iba z týchto dvoch atribútov sme vedeli predpovedať trhovú hodnotu hráča s chybou 1 milión.

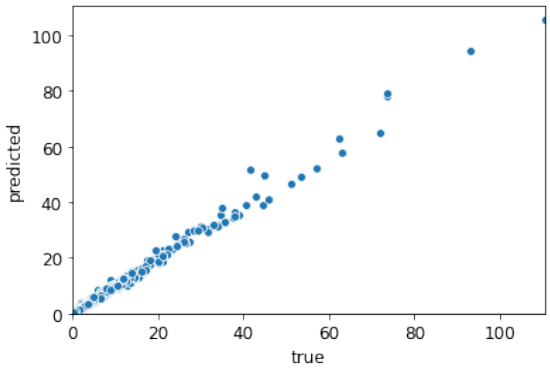


Fig. 9: Predikované hodnoty hráčov a reálne hodnoty hráčov (v miliónoch) našim najlepším modelom. (*RMSE = 0,474*).

11 Zhodnotenie

Oproti našim prvotným baseline modelom v sekciách 10.1 a 10.2 sa nám podarilo zlepšiť v oboch úlohách. V klasifikácií do trinástich pozícií sa nám nepodarilo dosiahnuť tak dobré výsledky ako pri klasifikácii do štyroch. Celkovo sa nám podarilo:

- natrénovať model na predikciu hodnoty hráča s *RMSE = 0,474* (v miliónoch)
- natrénovať model na predikciu hodnoty hráča iba z dvoch atribútov s *RMSE = 1* (v miliónoch)
- natrénovať model na klasifikáciu do 4 základných herných pozícií s úspešnosťami *0.896/0.906* (f1 micro/macro)
- so 100% úspešnosťou klasifikovať brankárov
- natrénovať modely iba zo schopností hráča a jeho charakteristík
- natrénovať modely s menej atribútami a s porovnateľnou úspešnosťou oproti použitiu všetkých atribútov

References

1. Alberto Fernandez, Salvador Garcia, Francisco Herrera, Nitesh V. Chawla: SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research **61**, 863–905 (2018), https://www.jair.org/index.php/jair/article/view/11192
2. Opitz, J., Burst, S.: Macro f1 and macro f1. arXiv preprint arXiv:1911.03347 (2019)
3. Razali, N., Mustapha, A., Yatim, F., Aziz, R.: Predicting player position for talent identification in association football. IOP Conference Series: Materials Science and Engineering **226**, 012087 (08 2017). https://doi.org/10.1088/1757-899X/226/1/012087
4. Soto-Valero, C.: A gaussian mixture clustering model for characterizing football players using the ea sports' fifa video game system. [modelo basado en agrupamiento de mixturas gaussianas para caracterizar futbolistas utilizando el sistema de videojuegos fifa de ea sports]. RICYDE. Revista Internacional de Ciencias del Deporte. doi:10.5232/ricyde **13**(49) (2017), https://www.cafyd.com/REVISTA/ojs/index.php/ricyde/article/view/1165
5. Yaldo, L., Shamir, L.: Computational estimation of football player wages. International Journal of Computer Science in Sport **16** (07 2017). https://doi.org/10.1515/ijcss-2017-0002
6. Zawbaa, H.M., Emary, E., Grosan, C., Snasel, V.: Large-dimensionality small-instance set feature selection: A hybrid bio-inspired heuristic approach. Swarm and Evolutionary Computation **42**, 29–42 (2018)