OZNAL - Dátová analýza dát hráčov z hry FIFA 19

Timotej Zaťko[1] and Tomáš Hoffer[2]

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

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

Abstract. Táto práca obsahuje analýza 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. Keďže dokázali vyjadriť tieto schopnosti hráčov číselne (na určitej škále), zaujíma nás ako, a či herná pozícia hráča závisí od týchto atribútov. Ďalšou zaujímavou otázkou je ako tieto atribúty (a ktoré najviac) vplývajú na trhovú cenu hráča v hre.

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.

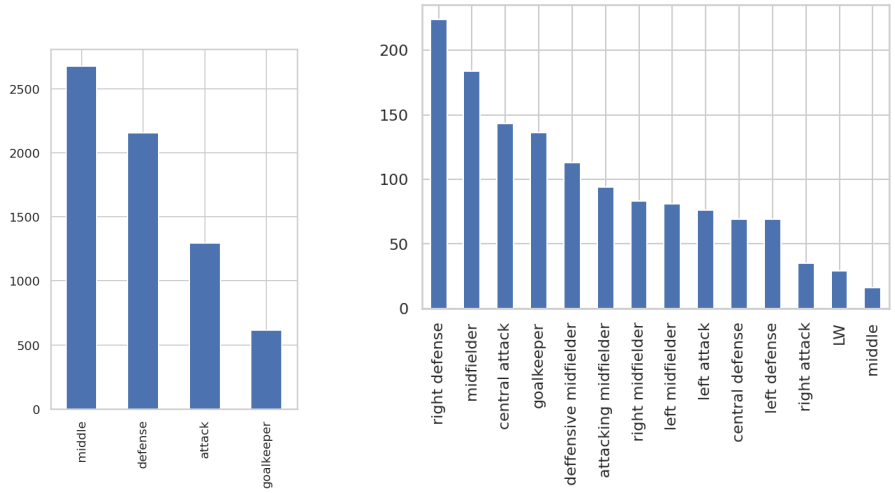
- **človek** - meno, národnosť, výška, hmotnosť, vek...
- **futbalový hráč (reálny svet)** - názov klubu, číslo dresu, dátum príchodu do klubu, dĺžka kontraktu, plat hráča, hosťovský klub, herná pozícia, preferovaná noha
- **futbalový hráč (hra)** - trhová hodnota hráča, triky, pracovitosť v útoku/obrane, 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

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. A 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 nejedná sa o chybu v úplnosti dátovej sady.

Analýzu sme 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' (Obr. 1). Môžeme pozorovať, že triedy nie sú vyvážené.

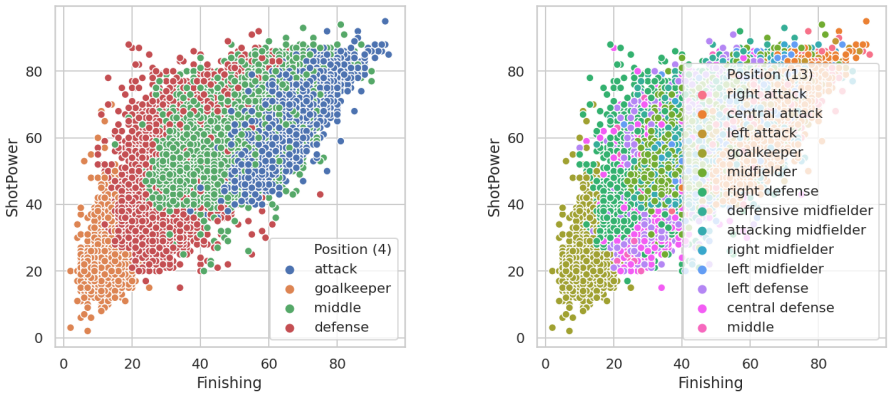
Pozíciu sa môžeme pokúsiť predikovať na základe 34 atribútov definujúcich futbalové schopnosti hráča určené hrou. Vizualizovali sme všetky dvojice týchto atribútov a pomocou "scatter plot" -ov sme hľadali zhluky jednotlivých pozícií. Ako príklad uvádzame vzťah medzi atribútmi 'Shot Power' (sila strely) a 'Finishing' (zakončovanie), kde môžeme pozorovať jednotlivé zhluky podľa pozície hráča (Obr. 2).



(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‘.



(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.

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 pravnú nohu, avšak na pozícii ľavého obrancu výrazne prevládajú ľaví obrancovia (Obr. 3).

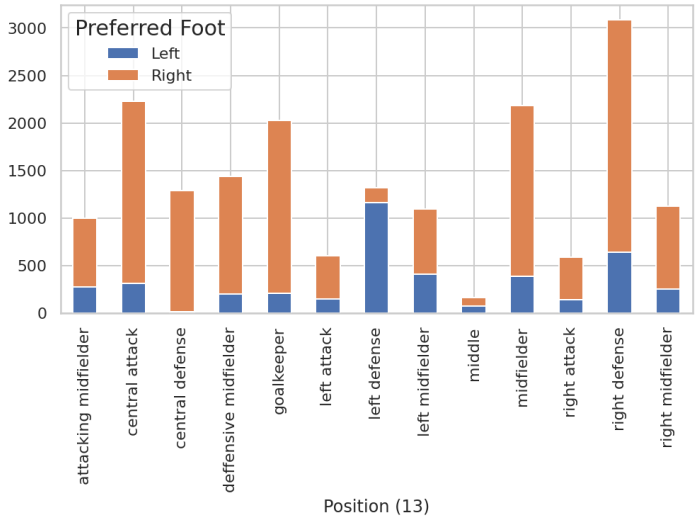


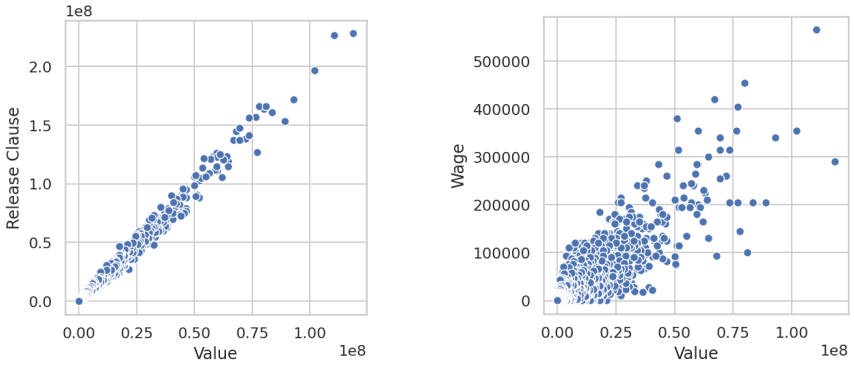
Fig. 3: Preferovaná noha hráčov podľa hernej pozície.

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

3 Definovanie úlohy objavovania znalostí

Rozhodli sme sa, že budeme vykonávať nasledovné úlohy:

– predikcia hernej pozície hráča – všeobecnej (4 triedy), rozšírenej (13 tried)



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

(b) Vzťah ‘Wage‘ voči ‘Value‘.

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

– predikcia hodnoty hráča (atribút s názvom ‘Value‘)

Obe tieto úlohy budeme realizovať z atribútov určujúcich herné shopnosti hráča a následne zo všetkých atribútov. Výsledky porovnáme, očakávame, že model natrénovaný zo všetkých atribútov bude výrazne lepší, 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
– odstránenie odľahlých pozorovaní
– 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.