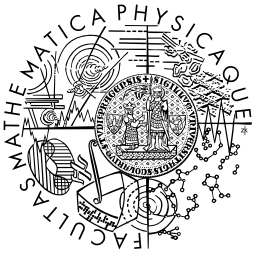


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Adam Kellich

Předpovídání výkonnosti fotbalových hráčů

Katedra teoretické informatiky a matematické logiky

Vedoucí bakalářské práce: Mgr. Martin Pilát, Ph.D.
Studijní program: Informatika
Studijní obor: Umělá inteligence

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Chtěl bych poděkovat všem lidem na fakultě, od kterých jsem se během uplynulých čtyř let mohl učit. Zejména děkuji vedoucímu práce Mgr. Pilátovi, Ph.D. za konzultace a cenné rady. To největší poděkování si ale zaslouží mí rodiče a prarodiče za neutuchající podporu, zájem na tom abych vystudoval a také za to, že když jim vyprávím, co jsem se na univerzitě naučil, tak předstírají, že mi rozumí.

Název práce: Předpovídání výkonnosti fotbalových hráčů

Autor: Adam Kellich

Katedra: Katedra teoretické informatiky a matematické logiky

Vedoucí bakalářské práce: Mgr. Martin Pilát, Ph.D., Katedra teoretické informatiky a matematické logiky

Abstrakt: Tato práce se zaměřuje na vývoj nástrojů pro zlepšení zážitku z hraní online fotbalové fantasy hry Sorare. Ve hře Sorare hráči kupují sběratelské kartičky, které reprezentují skutečné fotbalisty a soutěží s ostatními hráči, přičemž úspěch závisí na skutečných výkonech fotbalistů v reálných zápasech. Cílem naší práce je řešit dva hlavní problémy, kterým hráči ve hře čelí. Zaprvé se snažíme přesně předpovědět skóre fotbalisty v nadcházejícím zápase na základě údajů dostupných před zápasem. Zadruhé se snažíme identifikovat trhem podhodnocené hráče, což představuje potenciální investiční příležitosti. Popíšeme a implementujeme celý proces strojového učení od získávání dat, jejich zpracování, návrh vhodných algoritmů, trénování modelů až k vyhodnocení. V obou případech se navrhnuté algoritmy ukázaly jako užitečné oproti jednoduchým predikcím založených na průměru.

Klíčová slova: predikce časových řad strojové učení fantasy fotbal

Title: Football Player Performance Prediction

Author: Adam Kellich

Department: Department of Theoretical Computer Science and Mathematical Logic

Supervisor: Mgr. Martin Pilát, Ph.D., Department of Theoretical Computer Science and Mathematical Logic

Abstract: This thesis focuses on the development of tools to improve the experience of playing the online fantasy football game Sorare. In Sorare, players buy collectible cards that represent real footballers and compete against other players, with success depending on the actual performance of the footballers in real matches. Our work aims to address two main problems that players face in the game. First, we attempt to accurately predict a soccer player's score in an upcoming match based on data available before the match. Second, we seek to identify players undervalued by the market, which represent potential investment opportunities. We describe the whole process of machine learning from data acquisition, data processing, designing appropriate algorithms, training models to evaluation. In both cases, the proposed algorithms have demonstrated usefulness compared to simple average-based predictions.

Keywords: time-series prediction machine learning fantasy football

Obsah

Úvod

Fotbal, jakožto nejpopulárnější sport na světě, sledují každý týden miliony diváků. V případě Mistrovství světa ve fotbale 2022 v Kataru přesáhla celková sledovanost dokonce 1,5 miliardy.

V posledních letech se ve fotbale těšila velkému využití datová analýza (Pappalardo, 2019) a týmy jako Ajax Amsterdam, Brighton, RB Lipsko, RB Salzburg nebo Benfica Lisabon si v tomto odvětví udělaly jméno a dosáhly finančních úspěchů díky skautingu, rozvoji vyskautovaných hráčů a následnému prodeji. Ukázkovým příkladem může být obchod Ajaxu, který koupil tehdy dvacetiletého Brazilce Antonyho za 16 milionů eur a prodal v roce 2022 do Manchesteru United za 95 milionů eur.

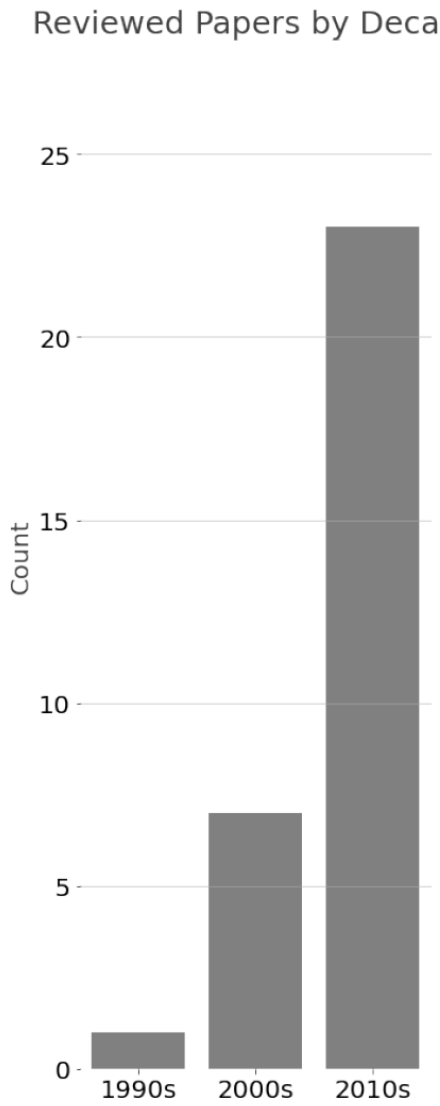
Soutěživost, sázkařství, fanouškovství nebo fotbalové fantasy přirozeně dávají za vznik snaze předpovídat výsledky zápasů, výkony hráčů, trendy trhu. Od roku 2010 se tato problematika dostala i zájmu vědeckých prací a v letech 2010-2020 bylo napsáno velké množství článků viz. obrázek 1 z (Bunker a Susnjak, 2022), který ilustruje, jak se v této oblasti od roku 2010 zintenzivnila činnost. Přes 100 článků zanalyzovali T. Horvat a J. Job ve své práci (Horvat a Job, 2020), která měla za cíl podat souhrn toho, co bylo v odvětví dosaženo. Výzkum je spjatý s rozvojem umělých neuronových sítí, které se ukázaly jako vhodný nástroj a které jsou nejpopulárnější (ne nutně nejlepší) využívanou metodou. Následované Naivní Bayesovskou klasifikací, SVM, lineární regresí a metodami založenými na rozhodovacích stromech.

Většina z těchto prací se zaměřuje na předpověď výsledku utkání, což je klasifikace do dvou nebo tří tříd (výhra, remíza, prohra – v některých studiích je vynechávaná remíza). V klasifikační úloze je nejlepším výsledkem reálného použití modelu na nadcházejících zápasech přesnost 82 %. Byl použit model natrénovaný na 4400 zápasech mezi lety 2002 a 2014 a vyhodnocený na 500 zápasech v sezóně 2016/2017 (Horvat a Job, 2020). Další zkoumané problémy jsou regrese pro předpověď gólového rozdílu mezi týmy nebo regrese pro předpověď skóre hráčů ve fantasy ligách. Tyto úlohy jsou považovány za náročnější než klasifikace z důvodu většího počtu možností (předpověď kdo vyhraje oproti tomu o kolik vyhraje) a většího vstupu náhody. U předpovědi skóre hráče zásadně záleží na tom, jestli mu bude z akce připsána asistence/gól, nebo dával pouze předfinální přihrávku a skóre se mu nezvedne. Může také být vystřídán, může mít špatný den, kdežto výsledek zápasu z pohledu jednoho týmu je závislý na 11 hráčích a výkonu v průběhu 90 minut.

V této práci se zabýváme využitím metod strojového učení v kontextu Sorare. Sorare (`www.sorare.com`) je fotbalová fantasy hra, kde hráči kupují NFT tokeny (s grafickým frontendem sběratelské karty) představující fotbalové hráče a následně z nich skládají týmy, prostřednictvím kterých spolu soutěží o ceny. Skóre týmů jsou přímo ovlivněny výkony hráčů v realitě. Nejpopulárnější fantasy hrou je dlouhodobě Fantasy Premier League (FPL), která se těší nejen největší hráčské základně, ale také největší podpoře pro výzkum v podobě datasetů, které komunita produkuje. Autoři práce Using ML Models to Predict Points in FPL (Bangdiwala a kol., 2022) predikovali body, které hráč ve fantasy zahraje a dosáhli úspěchu. Jejich RMSE (root mean square error) se pohybovala okolo 2.

Ve vlastní práci se zaměřujeme na dvě úlohy: predikce skóre hráče (regrese) v následujícím utkání v rámci Sorare na základě dat dostupných před zápasem; identifikace karet hráčů, kteří jsou podceněni trhem a můžeme u nich čekat nárůst ceny. K řešení těchto úloh jsme navrhli komplexní postupy zahrnující získání vhodných dat, jejich zpracování, výběr vhodných příznaků, což se ukázalo jako zásadní i se zjištěním, že větší dataset neznamená automaticky lepší predikce (Horvat a Job, 2020) a následné využití strojového učení k vytvoření informovaných odhadů.

Tato práce přispívá k rostoucímu počtu prací v oblasti fotbalových predikcí tím, že známou problematiku zkoumá v unikátním kontextu hry Sorare, což je kombinace, která ještě neexistuje. Velkým rozdílem mezi Sorare a FPL je ohodnocovací systém, který je v případě FPL méně složitý. Z tohoto důvodu není vhodné porovnávat výsledky studií z jiných fotbalových fantasy her s touto studií. Práce také přispívá tím, že porovnává efektivnost ML modelů před a po využití výběru příznaků, což není časté (Horvat a Job, 2020).



Obrázek 1: Počty prací na fotbalové téma v dekádách

1. Formální popis Sorare

Sorare je hra, ve které manažeři kupují karty, které následně používají ve fantasy. Každá karta má svou raritu a je spjata s jedním hráčem, který ovlivňuje její hodnotu na základě toho, jak se mu daří v jeho fotbalových zápasech. Manažeři své karty využívají v ligách a při dobrém umístění mohou získat odměny.

Tato kapitola definuje výše zmíněné pojmy a snaží se čtenáři přiblížit formát hry, reálie platformy (statistiky, ekonomiku) a také vysvětlit v jakých oblastech si slibujeme přínos od využití umělé inteligence.

1.1 Manažer

Manažerem rozumíme někoho, kdo si vytvoří Sorare účet, kupuje karty a soutěží. Tedy v přirozeném jazyce je to hráč Sorare, ale aby se to nepletlo s hráčem ve smyslu hráčem fotbalu, využíváme pojem manažer.

1.2 Hráč

Tímto pojmem se myslí reálný fotbalový hráč, jehož jménem je daná karta vytištěna. Některé soutěže mají omezení, že je možné použít jen hráče z určité ligy (např. Premier League). Pro využití karty ve fantasy ligách je rozhodující, na jakého hráče je karta vytištěna. Pokud hráč přestoupil do jiného klubu, karta je dále platná i přes to, že je vytištěná pro hráčův starý klub. Jeho skóre se vždy počítá ze zápasů za jeho momentální klub, nehledě na klub, za který hráč hrál v době vytištění karty.

1.3 Karta

Karta je ústředním pojmem Sorare. Jedná se o NFT token, který má i grafické rozhraní. Jedna karta je definována hráčem, raritou, sériovým číslem, sezónou, pozicí a úrovní. Karty do oběhu mezi hráče uvádí výhradně Sorare jejich vytisknutím a prostřednictvím aukcí (nejvyšší přihazující vyhrává) a až 33 % karet se k majitelům dostane prostřednictvím odměn za umístění v soutěžích. Existuje i sekundární trh mezi manažery, kde hráči směňují karty, které už v oběhu jsou prostřednictvím výměny za jiné karty, ETH (což je měna Sorare) nebo kombinaci obojího. Níže popisujeme jednotlivé atributy spjaté s kartami.

1.3.1 Rarita

Každá karta má svou raritu, kterou lze poznat jednak podle barvy a také podle čísla za lomítkem v seriovém čísle, které značí kolik karet existuje celkem. Tabulka 1.1 představuje rarity karet. Rarita karty také ovlivňuje její počáteční level, který dále ovlivňuje bonusová procenta, která karta obdrží ke svému skóre v daném herním týdnu. Limited začíná na levelu 0 a maximum je 20 (bonus od 0% do 10%). Rare začíná na levelu 0 a maximum je 20 (bonus od 0% do 10%).

Super Rare začíná na levelu 40 a maximum je 60 (bonusod 20% do 30%). Unique začíná na levelu 80 a maximum je 100 (bonus od 40% do 50%).

Rarita	Barva	Maximální počet za sezónu
Limited	žlutá	1000
Rare	červená	100
Super Rare	modrá	10
Unique	černá	1

Tabulka 1.1: Rarity karet

Obrázek 1.1: Příklad rare karty

1.3.2 Level

Level je součástí karty, ale ne jejího grafického rozhraní. Počítá se na pozadí a na základě levelu (který má rozsah od 0 do 20) získává karta procentuální bonus, který zvyšuje počet bodů, které karta obdrží v soutěžích. Level se zvyšuje se získaným počtem XP. XP jde získat využíváním karty v soutěžích.

1.3.3 Sezóna

Sezóna značí, ve které fotbalové sezóně byla karta vydána. Evropský formát pro sezónu je (rok-1)/rok (např. 2021/2022), protože sezóna začíná v létě a končí v létě příštího roku. Pro karty hráčů z asijských/amerických klubů se využívá jen rok (např. 2021), protože sezóna tam začíná na začátku a končí na konci roku.

Důležité je, že pokud je pro danou kartu sezóna maximální (neexistuje karta téhož hráče, kde by sezóna byla vyšší - tedy aktuálnější), získává karta další bonus 5 %.

1.3.4 Sériové číslo

Sériové číslo je identifikátor karty hráče v rámci jedné sezóny. V jedné sezóně existuje pouze jedna karta $n/100$ pro jednoho hráče. Dá se spekulovat, že karty, kde se sériové číslo shoduje s číslem hráče na dresu, jsou hodnotnější. Obecně platí, že karty s nižším n jsou v aukcích dražší z psychologického důvodu. Design karty je v tu chvíli nový, nejdéle u ní bude platit bonus 5 % za aktualitu a často i pro daného hráče zatím neexistovaly na Sorare karty. S rostoucím n zájem klesá a s tím i cena v aukcích.

1.3.5 Pozice

Pozice, na které lze hráče využít v rámci lineupu nabývá hodnot buď brankář, obránce, záložník, útočník (goalkeeper, defender, midfielder, forward). Existují hráči u kterých se pozice měnila a např. Ferran Torres z Barcelony má karty s pozicí midfielder i forward. Z důvodu odlišností mezi jednotlivými pozicemi se v práci ve všech ohledech (regrese, clusterování...) přistupuje ke každé pozici zvlášť.

1.4 Sestava

Sestava (lineup) je uspořádaná čtveřice hráčů. Na první pozici musí být goalkeeper, na druhé defender, na třetí midfielder, na čtvrté forward. Ve většině soutěží je součástí sestavy i extra pátý hráč, který může být libovolné pozice. Sestava musí také respektovat restrikce té soutěže, do které ji chceme registrovat. V rámci sestavy často manažer volí i kapitána - kartu, která obdrží navíc 20/30/50 procent bodů, dle pravidel soutěže.

1.5 Soutěž

Jsou dvě základní dělení soutěží – podle rarity a podle restrikcí na karty, které manažeři smí používat. Dělení podle rarity znamená, že v soutěži Rare Champion Europe lze používat jen červené karty. Champion Europe znamená, že v této lize smí manažeři využívat pouze hráče z top 5 evropských lig (německá, španělská, anglická, italská, francouzská) a jedná se tedy o restrikci. V době psaní práce došlo ke změnám v soutěžích a přibyly zajímavé formáty, které se jmenují Cap 220, 240, 270. Pro Cap 220 platí, že součet všech průměrů hráčů za posledních 15 zápasů (L15) nesmí překročit 220, analogicky pro ostatní. S touto formulací soutěží se nabízí, po natrénování modelů pro predikci skóre, využít programování s omezujícími podmínkami nebo evoluční algoritmy pro nalezení sestavy s nejvyšším predikovaným skóre, která zároveň splňuje podmínky soutěže.

1.6 Herní týden

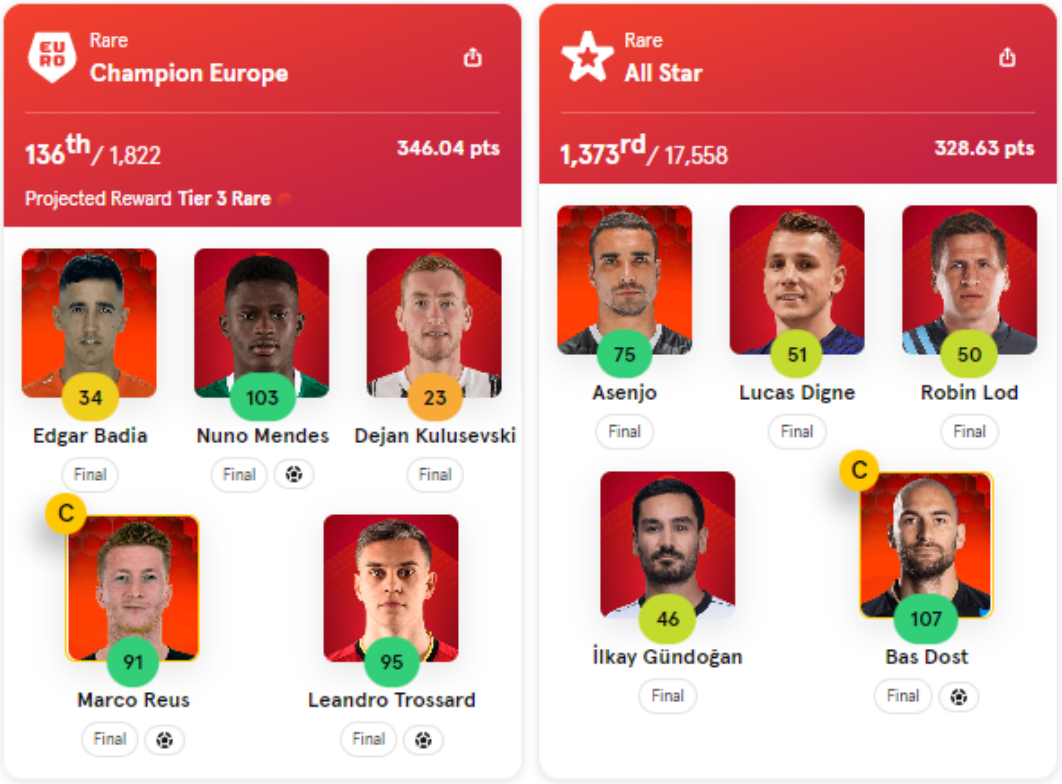
Soutěžení v Sorare je uspořádáno do jednotlivých hracích týdnů, anglicky gameweek. Jedná se o vymezený časový úsek, typicky trvající 3 až 4 dny, ve kterém mohou manažeři získávat body v soutěžích. Mechanismus získávání bodů je popsaný více v kapitole 1.6.1. Za tyto body se manažeři umístí v žebříčku a na základě umístění získávají odměny. Potřebné body a následné odměny jsou popsány v podsekci 1.6.2

1.6.1 Jak se vypočítá skóre hráče

Skóre hráče se skládá ze dvou složek a to „decisiveScore“ a „allAroundScore“ a výsledné skóre je součtem těchto dvou složek. Decisive score je ovlivněno akcemi, které hru mění nejvíc. Pozitivní akce jsou: vstřelení gólu, asistence na gól, vyhraná penalta, vykopnutí míče z brankové čáry, čisté konto a chycená penalta (pro brankáře), obraný zákrok jakožto poslední hráč. Negativní jsou: vstřelení vlastního gólu, červená karta, faul, kvůli kterému soupeř kope penaltu, chyba zapříčiňující gól, obdrženo 3 a více gólů (pro brankáře). Pozitivní/negativní akce zvedne decisive score o jednu úroveň výše/níže. Každá úroveň má svůj přidělený počet bodů (0, 5, 15, 35, 60, 70, 80, 90, 100), kde startovní hodnotou pro hráče, který zápas začíná v základní sestavě, je 35. *AllAroundScore* je ovlivněno statistikami, které hráč za utkání nasbírá. Přepočet statistik na *AllAroundScore* je skalární součin vektoru statistik a vektoru ze scoring matrix 1.2 pro danou pozici hráče. Výsledné skóre je ještě přenásobeno číslem: 1 + bonusová procenta karty, toto bonusové přenásobení ale neplatí pro Cap soutěže.

1.6.2 Umístění hráčů a odměny

Skóre je vypočítáno pro všechny hráče a celkové skóre sestavy je potom suma přes všechny hráče v něm. Před startem herního týdne je vždy dopředu vypsaný prizepool, který sestává z peněžních cen pro nejpřednější místa a z karet pro ostatní. Prizepooly se dynamicky mění podle počtu hráčů, ale složitost hry se dá zhruba popsat následovně. První (nejslabší) ceny získává manažer, když se umístí alespoň zhruba v horních 5 procentech. Výjimkou jsou soutěže s názvem Cap 240 (dříve All Star), ve kterých jsou aktivní tzv. threshholdy. Pokud manažer nahraje alespoň 250 bodů v limited/rare/super rare/unique All Star získává obnos v ETH v té chvíli odpovídající 5$/50$/200$/500$.



Obrázek 1.2: Příklady lineupů po zkompletovaném gameweeku

Action	Goalkeeper	Defender	Midfielder	Forward
General				
Was Fouled	0	0	1	1
Fouls	-1	-2	-1	-0.5
Error Leads to Shot	-5	-5	-3	-3
Yellow Card	-3	-3	-3	-3
Defending				
Goals Conceded	-3	-4	-2	0
Clean Sheet	0	10	0	0
Effective Clearance	0	0.5	0	0
Tackle won	0	3	3	0
Blocked cross	0	1	1	0
Block	0	2	1	0
Double Double	0	4	4	4
Double Triple	0	6	6	6
Triple Triple	0	12	12	12
Possession				
Possession Won	0	0.5	0.5	0
Possession Lost	-0.3	-0.6	-0.5	-0.1
Duel Won	0	1.5	0.8	1
Duel Lost	0	-2	-0.8	-1
Interception	0	3	3	3
Passing				
Big Chance Created	3	3	3	3
Key Pass	2	3	2	2
Accurate Pass	0.1	0.08	0.1	0.1
Accurate Final Third Pass	0.5	0.4	0.3	0.1
Accurate Long Ball	0.2	0.5	0.5	0
Long Pass into Opposition Half	0	0.5	1	0
Missed Pass	0	-0.2	-0.3	0
Attacking				
Shot on Target	3	3	3	3
Contest Won	0	0.5	0.5	0.5
Penalty Area Entry	0	0.5	0.5	0.5
Penalty Missed	-5	-5	-5	-5
Big Chance Missed	-5	-5	-5	-5
Goalkeeping				
Inside Box Save	2	0	0	0
High Claim	1.5	0	0	0
Punches	1.5	0	0	0
Diving Save	3	0	0	0
Diving Catch	3.5	0	0	0
Unclaimed Cross	-3	0	0	0
Six Second Violation	-5	0	0	0
GK Smother	5	0	0	0
Keeper Sweeper	3	0	0	0

Tabulka 1.2: Scoring matrix

2. Zdroje dat

Jako primární zdroj dat využíváme API poskytnuté samotnou platformou Sorare. V těchto datech, z důvodu toho, že jsou využívána pouze v rámci Sorare hry, chybí zásadní informace pro jakoukoliv fotbalovou analýzu (např. cena hráče na reálném trhu). Je nutnost tato data obohatit z dalších zdrojů. Pro obohacení dat využíváme stažení z webu či vlastní implementace webscrapingových technik. V této kapitole popíšeme, jaké datové zdroje jsme zvolili a jak jsme data získali. Na příkladech vysvětlíme, jaký typ příznaků každá datová sada obsahuje.

2.1 Sorare API

Sorare API využívá dotazovací jazyk GraphQL.
Pro napsání programu, který zajistí stažení dat ze SorareAPI jsme zvolili C# z důvodu dobré podpory pro GraphQL (knihovna GraphQL). Program se připojí k API a následně posílá dotazy a přijatá data ukládá typicky ve formátu json. Hlavními datonosnými soubory, které program vytvoří jsou all_players.json a all_cards.json. Na tomto odkaze si lze vyzkoušet níže přiložené dotazy: `https://api.sorare.com/graphql/playground`.

V obrázku 2.1 je vložen GraphQL dotaz, který stahuje data pro hráče "Lionel Messi"a v obrázku 2.2 dotaz, který stahuje první 2 karty uložené v databázi Sorare. Stahuje karty libovolné rarity a musí to být karty, které už jsou v oběhu (owned je true). Pro stažení všech karet na platformě bylo potřeba využít graphQL kurzor, který ukazuje na poslední staženou kartu a pomocí něj se vždy dostat k další stránce lineárního spojového seznamu allCards.

2.1.1 All players

Data stažená ze Sorare API strukturálně odpovídají GraphQL dotazu 2.1, jen jsme pro práci rozšířili množinu dotazovaných atributů. Stažený json soubor jsme zpracovali a výsledkem jsou dva csv soubory, all_players_simple.csv a game_stats.csv. V prvním jmenovaném souboru odpovídá každý řádek jednomu hráči a informacím, které o něm platí v moment stažení dat: věk, pozice, v jakých klubech hrál (clubMemberships), kolik karet je vedeno pod jeho jménem na Sorare (cardSupply), či například jaký je jeho průměr skóre z posledních 5 zápasů (lastFiveSo5AverageScore). Počet hráčů je 20 až 25 tisíc.

Soubor, který dále (v práci i v kódu) nazýváme game_stats, vznikl zpracováním atributu "allSo5Scores", což je seznam zápasů a k nim příslušných informací. Každý řádek v game_stats je 1 fotbalové utkání z pohledu jednoho hráče. Sloupce jsou statistiky, které byly pro hráče zaznamenané v daném zápase. Množina statistik je popsaná v tabulce 1.2. Jeden ze sloupců je "totalScore", což je cílová veličina v kapitole 3, a proto každý řádek představuje jeden potenciální datový záznam pro učení s učitelem.

Celkem je v game_stats k dispozici okolo 65 příznaků a počet zápasů je skoro 800 tisíc. Jedná se ale o nekvalitní data, jejichž zpracování bude popsáno v dalších kapitolách. Jsou nekvalitní, protože většina zápasů z pohledu hráče je takových, že

```
query statsForOnePlayer {
player(slug:"lionel-andres-messi-cuccittini") {
        slug
    clubMemberships{club{slug} startDate}
    position
    age
    appearances
    status{
      lastFiveSo5Appearances
      lastFiveSo5AverageScore
      lastFifteenSo5Appearances
      lastFifteenSo5AverageScore
    }
    country{code}
    cardSupply{limited rare superRare unique}
         allSo5Scores(first:1){nodes{
                game {date}
                score
                decisiveScore {totalScore}
                  positiveDecisiveStats {category stat statValue}
                  negativeDecisiveStats {category stat statValue}
                allAroundStats {category stat statValue}
    }}
}}
```

Obrázek 2.1: All players query.

```
query allCards {
  allCards (
   rarities:[limited,rare,super_rare,unique] owned:true first:2
   )
   {
    nodes {
        slug
        player { slug }
        powerBreakdown { season xp scarcity }
        rarity
        tradeableStatus
        serialNumber
        latestEnglishAuction {
          startDate
          bidsCount
          bestBid { amount amountInFiat { eur } }
        }
        notContractOwners { from price transferType }
      }
    }
}
```

Obrázek 2.2: All cards query.

nehraje a tedy má všechny statistiky nulové. Takový řádek nese nulovou informaci. Také jsou častá chybějící data v jednotlivých sloupcích.

2.1.2 All cards

All cards jsou data stažená ze Sorare API a strukturálně odpovídají GraphQL dotazu 2.2, jen jsme pro práci rozšířili množinu dotazovaných atributů. Stažený json soubor jsme zpracovali a výsledkem je jeden csv soubor all_transactions.csv. V tomto souboru jeden řádek odpovídá jedné transakci (karta mění majitele), o které víme cenu v ETH, cenu v eurech, majitele, informace o aukci, která uvedla kartu do hry a další.

2.2 Football data

Přímo byla stažena data z `https://www.football-data.co.uk` a v kódu i v práci jsou dále označována jako league_stats. Jedná se o často používaná data pro podobné analýzy Horvat a Job (2020). Pro každou ligu máme dlouhou (často zpět až do roku 2012) historii všech zápasů, které se v rámci ní konaly. Jedná se o obecné statistiky popisující utkání jako: výsledek, počet gólů domácího/hostujícího týmu, ale co je nejdůležitější - kurzy vypsané sázkovými kancelářemi. Kurzy pro výhru obou týmů, remízu, že bude v zápase počet gólů větší/menší než 2.5, asijský handicap na domácí apod. Pro známější evropské

soutěže je k dispozici 106 příznaků, dále nazýváme tyto tabulky jako typ "euro". Pro méně známe evropské, či neevropské soutěže 19 příznaků, dále jako "world". V obou typech jsou příznaky, které jsou spočítany jako agregační funkce přes kurzy v jednotlivých monitorovaných sázkových kancelářích. Příkladem může být dvojice MaxA, AvgA, které značí jak vysoký byl maximální kurz na výhru hostů a jaký byl průměrný kurz na výhru hostů. Ve world tabulkách je obsaženo to nejpodstatnější a to, co mají tabulky euro navíc jsou typicky pouze kurzy ve větším počtu konkrétních sázkových kancelářích, ze kterých se agregace počítala. Příznaky, která později využívám, jsou zpravidla ty, které jsou vyrobeny agregačními funkcemi.

2.3 Understat

Jedná se o data dostupná na `https://www.understat.com`. Na understat jsou data jen pro tyto ligy: "ligue-1-fr", "serie-a-it", "laliga-santander", "premier-league-gb-en", "bundesliga-de", "russian-premier-league". Nejde je přímo stáhnout, tak využíváme vlastní python script s pomocí tohoto balíčku `https://understat.readthedocs.io/en/latest/`, který se k webové stránce připojí a v námi zvoleném formátu je stáhne. Hlavním důvodem, proč jsme stahovali data z tohoto zdroje jsou statistiky typu xG, xA, xPTS (expected points). Malé x v tomto případě znamená expected a veškeré statistiky začínající na x výchazejí z xG - expected goals.
Expected goals xG je relativně nový druh statistiky, který se začal používat až v posledních letech. Jedná se o pravděpodobnost, že střela z dané pozice skončí gólem. Když je pro pozici xG 0.2, očekáváme, že hráči z 10 takových šancí dají 2 góly. xG statistiky si vytváří firmy zabývající se danou problematikou, které mají k dispozici dostatek historických dat, na jejichž základě je počítají. Pro každou šanci bere v úvahu např. vzdálenost od branky, úhel k brance, část těla, kterou hráč střílel, typ přihrávky, kterou dostal apod. xG jako statistika vznikla z potřeby lepšího vyjádření toho, jaký tým má v zápase převahu. Do té doby nejlepší způsob bylo podívat se na držení míče nebo střely na bránu. Konkrétní stažená data popíšeme v následujích podsekcích.

2.3.1 league_tables

Jako league_tables značíme tabulky pro sledované soutěže pro sezóny 2014 a aktuálnější. Tabulky jsou vždy ve stavu dohrané ligy, kromě těch aktuálních, kde stav ligy odpovídá datu stažení. V příznacích se dá najít počet odehraných zápasů, počet výher/remíz/proher, počet vstřelených/obdržených gólů a další statistiky jako například kolik tým dovolí soupeři přihrávek na své polovině (PPDA). Jeden řádek v jedné konkrétní tabulce odpovídá tomu, jak si tým v daný moment v lize vede.

2.3.2 league_stats

League_stats jsou statistiky indexované opět pomocí dvojice (liga, sezóna), které popisují počet vstřelených gólů a očekávaných golů pro domácí/hostující

týmy. Z toho se dá odvodit například ofenzivnost ligy nebo výhodu jakou poskytuje domácí prostředí.

2.3.3 team_stats

Team_stats jsou statistiky indexované pomocí dvojice (tým, sezóna), které popisují do detailu, jak se tým chová v při různých stavech zápasu, nebo z jakých situací dává góly. Stavy jsou například goalDiff+1, goalDiff-1 (vedení, prohrávání týmu o 1 gól), timing 1-15 (události v prvních patnácti minutách). Situace jsou openPlay, fromCorner... Plným příkladem statistiky je pak např. goalDiff+1 Goals, což značí kolik dá tým gólů, když vede o 1 gól.

2.3.4 players

Players značíme tabulku indexovanou pomocí dvojice (hráč, sezóna). Dostupné jsou převážně ofenzivní statistiky jako g, xg, a, xa, střely, klíčové přihrávky, xGChain - definované jako xG akcí, kterých se hráč účastnil. xGChain mají vyšší ti hráči, kteří se účastní nebezpečných situací u soupeřovo brány a je u nich tedy větší šance, že dosáhnou vysokého decisiveScore. xGBuildup je podobné, ale z těchto akcí jsou vyjmuty finální přihrávky a střely. Vysoký xGBuildup mají typicky hráči, kteří hrají ve středu pole, jsou součástí většiny akcí, ale málokdy se dostávají do zakončení.

2.4 FIFA

Hra FIFA je vhodným zdrojem dat, protože do ohodnocování hráčů jde každý rok, před vydáním nové verze FIFY, hodně práce expertních týmů. Výsledkem je tabulka, kde jeden řádek obsahuje popis hráče napříč statistikami pomocí skóre 0 až 100. Hlavní statistikou je celkové skóre, kde nejlepší hráči v top týmech dosahují maxima 94 (Lionel Messi) a hráči v slabších ligách mají skóre okolo 60. Nejlepší čeští hráči (hrající např. v Premier League) mají skóre okolo 75. Příkladem ofenzivních statistik je střelba, zakončení, přihrávky. Příkladem defenznivních statistik je bránění, schopnost skluzování, hlavičkování. Mimo tyto statistiky lze v řádku najít třeba odhadnutí potenciálu hráče, hráčovu reálnou peněžní hodnotu, jak umí střílet slabší nohou apod. Celkem je k dispozici záznam pro 92000 hráčů a 111 příznaků.

3. Predikce výkonu hráče

Predikce výkonu hráče je problém, kde na základě dat, které máme v daný časový moment k dispozici, chceme predikovat číslo 0-100 značené *totalScore*. Zvolili jsme postup ručního vygenerování velké datové sady (feature_pool) ze všech datových zdrojů a následným automatickým výběrem příznaků. V tomto případě je autorský zásah do finální datové sady použité pro strojové učení mimimální a to pouze ve fázi přípravy feature_pool. Tento automatický výběr bychom chtěli porovnat s ručním výběrem, kdy vybereme vhodnou podmnožinu příznaků, čistě na základě fotbalových znalostí a vlastního úsudku.

Myšlenkou automatizovaného přístupu tedy je zpracovat datové zdroje a vytvořit z nich datovou sadu, které říkáme **feature_pool** a která má za cíl obsáhnout co nejvíce informace, byť přebytečné. Tu použijeme pro natrénování modelů a vyhodnotíme. Hypotéza je, že tento přístup pravděpodobně není optimální, protože při velkém počtu příznaků dochází k prokletí dimenzionality a k přeučení. Autoři často preferují nižší počet příznaků a dobré výsledky v přechozích studiích byly v nepřímé úměře s počtem příznaků Horvat a Job (2020). Z těchto důvodu bude dalším důležitým krokem feature engineering a feature selection.

Metody pro výběr příznaků a jiné myšlenky popsané níže dávají za vznik několika parametrům pro strojové učení, ze kterých byly vytvořeny různé kombinace a celkem bylo natrénováno přes 700 modelů. Výsledky rozebíráme v kapitole 5.

3.1 Vytváření příznaků: feature_pool

Zpracovávali jsem všechny 4 datové zdroje individuálně a výsledky spojili do jedné tabulky **feature_pool**. Tento přístup má výhodu v jednoduchosti a intepretabilitě výsledného datasetu, kde při výběru příznaků můžeme říct, který datový zdroj poskytl nejlepší informaci. Za zvážení by stálo i kombinovat informace z různých datových zdrojů a vytvářet z nich příznaky, jako například xG hráče děleno velikostí jeho střeleckého atributu ve FIFA, což by dalo za vznik indikátoru, který by říkal, jestli hráč v dané sezóně dostává svých kvalit.

Níže pro každý datový zdroj popisujeme, jak pro jeden konkrétní zápas jednoho hráče a odpovídající *totalScore* vytvoříme příznaky, které toto skóre mohly ovlivnit.

1. **Pro game_stats** nejdříve vyfiltrujeme zápasy, kde hráč hrál méně jak 60 minut. Pro hráče kteří hrají menší počet minut vstupuje do hry více náhody a je těžší cokoliv predikovat. Také se ve vyloučené množině nachází hráči, kteří neodehráli ani minutu a cílem práce není se pokoušet predikovat, jestli hráč nastoupí, nebo ne. Předpokládáme, že hráč pro kterého bude model predikovat, hrát bude.

   Pro jeden původní game_stats řádek si nejdříve zjistíme všechny spoluhráče hráče a všechny protihráče a pomocí nich získáme tabulky, které obsahují jen zápasy těchto hráčů. Také samozřejmě využíváme tabulku se zápasy pro konkrétního hráče. Na těchto tabulkách spočítáme průměr a směrodatnou odchylku pro každou statistiku pro posledních n zápasů, kde n jsme volili

40, čímž dostáváme statistiky typu "statname_L40". Ty se dají interpretovat např. takto: totalScore_L40 = jaké měli průměrné skóre spoluhráči daného hráče, enemy_shots_L40 = jaký počet střel průměrně vyprodukují protihráči, kolik faulů udělají protihráči apod. Vytvoříme také binární indikátor, jestli byl zápas v domácím prostředí nebo v hostujícím. Příznaky z této kategorie značíme předponou gs jako game_stats.

2. **Pro league_stats** je hlavním cílem popsat kontext a sílu soupeřova i hráčova týmu na základě statistik a kurzů sázkových kanceláří. Zde jsme vybrali podmnožinu statistik, které nás zajímají a typicky jsou spočítány jako průměr/maximum přes všechny vypsané kurzy na danou událost. Příkladem jsou 'goalsScored', 'goalsConceded', 'avgWinOdds', 'avgDrawOdds', 'maxWinOdds', 'maxDrawOdds', 'shotsOnTarget', 'shotsOnAgainst', 'shots', 'shotsAgainst', 'avgUnder2.5', 'avgOver2.5', 'maxUnder2.5', 'maxOver2.5', 'avgCAH', 'avgCAHAgainst', 'maxCAH', 'maxCAHAgainst'.

   Statistiky obsahující text "AH"znamenají asijský handicap, tedy že tým vyhraje o více než jeden gól. Over a under 2.5 pak znamená, jestli v zápase bude více nebo méně gólů než 2,5. Nejdříve si vyhledáme všechny zápasy hráčova i soupeřova týmu a pro oba vyrobíme průměrné statistiky za posledních 5, 15, 40 zápasů, což může pomáhat se zachycením trendu, jakým se ubírá výkonnost mužstva v čase. Dále byly statistky rozděleny na domácí/hostující/všechny, protože z Rodrigues a Ângelo Pinto (2022) vyplývá, že prostředí výrazně ovlivňuje zápas. V domácím prostředí týmy vyhrají 45% zápasů a tedy poskytuje velkou výhodu. Některé týmy jsou na domácí prostředí citlivější než ostatní (např. pověstná pevnost Liverpoolu na Anfield Row). Příznaky z této kategorie značíme předponou LS.

3. **Pro data z understat** jsme se rozhodli sledovat vždy nejaktuálnější dohranou sezónu, aby nedošlo k úniku dat (trénování na datech, která v daný moment nemohla být dostupná). Šlo by k učení použít data z aktuální sezóny a zejména v případě, že už je odehráno hodně zápasů, by to určitě vylepšilo výsledky, ale struktura understat toto lehce neumožňuje. Protože data ze Sorare API používají jiná jména pro týmy a hráče, bylo nutno tato jména automaticky napárovat pomocí python knihovny fuzzywuzzy a výsledky manuálně doupravit, pokud párování někde proběhlo špatně (například inter-milano-milano se vždy mapuje na Milano, ale to je AC Milán, ne Inter Milán). Nutnost párovat jména platí i pro data league_stats a z FIFA. Data z tohoto zdroje jsou vhodná k užití tak, jak jsou, protože popisují hráče/tým/ligu v daný moment, což je přesně to, co potřebujeme. Z toho důvodu je do feature_pool vložíme bez dalšího zpracování. Data poskytují 4 druhy informací: statistiky hráčova/soupeřova klubu v dané sezóně, statistiky hráče v dané sezóně - rozšířeno o statistiky soupeřových hráčů a statistiky spoluhráčů, statistiky ligy v dané sezóně a deskriptivní statistiky o tom, jak se chovají oba týmy za určitých stavů zápasu v dané sezóně. Příznaky z této kategorie značíme předponou u.

4. **Pro data z FIFA** byly zjištěny všechny roky, kdy se daný hráč ve FIFA objevil a byl vybrán ten nejrelevantnější (menší roven roku konání zápasu a zároveň mu nejblíže, např. pro zápas v květnu 2020 vybereme data z FIFA

19). Stejným způsobem byly identifikovány množiny všech spoluhráčů a protihráčů a z nich spočítán průměr přes jednotlivé statistiky, pro popis průměrné kvality soupeřova a hráčova týmu. Takto vybraná data jsou vhodná do feature pool, protože popisují hráče, hráče soupeře a spoluhráče dlouhodobě a kvalitativně na základě všeho, co se o nich vědělo před sezónou. Příznaky z této kategorie značíme předponou f.

3.2 Výběr příznaků

Celkově má vygenerovaný feature_pool okolo 1300 příznaků a tak je tento krok výběru velice důležitý. Existuje velké množství způsobů, jak příznaky vybírat, z nichž některé jsou výpočetně velice náročné (iterativní výběr). Uvádíme tři techniky, které jsme využili a čtvrtá je ruční výběr, který na nich staví.

1. **Na základě korelací:** Vypočítáme si pro každý příznak z feature_pool její korelaci s cílovou veličinou *totalScore*. Do finální datové sady vybíráme pak pouze ty, které mají absolutní hodnotu větší než zadaný práh th, a nebo podle parametru n který určí, že vybereme n příznaků s největší absolutní hodnotou korelací.

2. **Na základě random forest:** Na největším dostupné datové sadě (feature_pool s 1300 příznaky) natrénujeme model RandomForest a využijeme parametr feature importance k výběru nejlepších n příznaků, nebo příznaků, které mají větší důležitost než zadaný práh th. Práh z implementačních a dokumentačních důvodů používáme ze stejné škály jako jsou korelace (tedy 0 až 1 v absolutní hodnotě), ale pro rf selekci ho dělíme 100. Je to z důvodu, že hodnoty feature importance se vždy nasčítají na 1 a příznaků v původní sadě bylo přes 1000 a hodnoty jsou tedy malé. Proto když je v tabulce 5.1 ve vyhodnocení uvedena kombinace rf_0.15, znamená to, že byly využity příznaky s hodnotou feature importance 0,0015 a větší.

3. **Na základě PCA:** PCA (Principal component analysis) je technika pro redukci dimenzionality vstupu, široce využívaná ve strojovém učení. Funguje na principu maximalizace rozptylu a vstupní data zobrazuje do méně dimenzionálního prostoru. Její síla spočívá zejména v dobrém reprezentování obrazových dat. V mém případě, pro fotbalová data, nečekám, že bude fungovat až tak dobře.

4. **Ruční selekce:** První dvě použité selekce nám poskytují seřazení příznaků podle definované důležitosti. Toto seřazení bylo bráno v potaz a s myšlenkou dostatečného pokrytí a zároveň minimalizace přebytečné informace byla vybrána podmnožina příznaků pro každou pozici.

3.3 Použité modely

Na základě nastudované literatury (Horvat a Job, 2020) jsme zvolili 6 standardně využívaných modelů vhodných k vyzkoušení: Multi layer perceptron (MLP), Support vector regressor (SVR), Linear regression (LR), Gradient boosting regressor (GBR), Random forest (RF), XGBoost regressor (XGBR). Využili jsme

převážně implementací z balíčků sklearn a xgboost. K nastavení hyperparametrů není vhodné používat cross-validaci z důvodu chronologické povahy dat. K nastavení hyperparametrů byla ve finální verzi použita validační datová množina a framework Optuna.

3.3.1 MLP - sklearn.neural_network.MLPRegressor

MLP je typ umělé neuronové sítě, která se skládá z více vrstev vzájemně propojených neuronů. K trénování používá zpětné šíření, které minimalizuje ztrátovou funkci upravováním matic vah. MLP často používají jako ztrátovou funkci střední kvadratickou chybu (MSE). Mezi klíčové hyperparametry pro MLP patří počet skrytých vrstev, počet neuronů v jednotlivých vrstvách, aktivační funkce, rychlost učení a regularizační konstanta. Aktivační funkce je nelineární funkce, která se používá na výstup neuronů ve skrytých vrstvách a umožňuje tak neuronové síti řešit nelineární a komplexní problémy.

3.3.2 SVR - sklearn.svm.SVR

Support Vector Regressor (SVR) je model, který se zaměřuje na predikci spojitých hodnot nalezením optimální nadroviny, která odpovídá vstupní datům. Cílem SVR je minimalizovat chybu mezi předpokládanými a skutečnými hodnotami a zároveň zachovat specifickou rezervu. Toho je dosaženo zavedením oblasti (margin) s pevnou velikostí kolem regresní nadroviny, kde se žádná data nacházet nesmějí, jinak jsou ve ztrátové funkci penalizována. K hyperparametrům patří regularizační konstanta C, která zajišťuje míru kompromisu mezi minimalizací odchylky predikcí od nadroviny a penalizací za to, že se datový bod octne v marginu. SVR díky tomuto principu dobře zobecňuje na neviděná data a funguje dobře jak při nízkém počtu dat, tak příznaků. Naproti tomu je citlivý na škálování a vzájemně korelované příznaky.

3.3.3 LR - sklearn.linear_model.Ridge

Lineární regrese (LR) je jednoduchý statistický model, který předpokládá lineární vztah mezi vstupními příznaky a cílovou veličinou. Jeho cílem je najít optimální váhy pro každý příznak minimalizací MSE. LR má pouze regularizační hyperparametr alpha díky kterému je méně náchylný k přeučení. LR funguje dobře na datech s lineárními vztahy, takže za předpokladu dobrého výběru příznaků je vhodný pro fotbalovou doménu.

3.3.4 RF - sklearn.ensemble.RandomForestRegressor

Random forest (RF) je metoda, která konstruuje více rozhodovacích stromů a kombinuje jejich předpovědi. RF je odolný proti přeučení a dobře zvládá korelované příznaky díky náhodnému výběru a technikám bootstrappingu. Mezi klíčové hyperparametry RF patří počet stromů, maximální hloubka a počet rysů uvažovaných při každém rozdělení. RF funguje dobře na široké škále dat, včetně těch vysokodimenzionálích se složitými vztahy.

3.3.5 GBR - sklearn.ensemble.GradientBoostingRegressor

Gradient Boosting Regressor (GBR) je metoda strojového učení, která postupně konstruuje několik rozhodovacích stromů, kde každý strom se snaží opravit chyby, kterých se dopustil předchozí strom. Používá gradient descent k minimalizaci ztrátové funkce (např. střední kvadratické chyby). Mezi klíčové hyperparametry GBR patří počet stromů, rychlost učení, maximální hloubka stromu a ztrátová funkce. GBR funguje dobře na různých souborech dat, ale může být citlivý na šum a korelované rysy.

3.3.6 XGBR - xgboost.XGBRegressor

Lepší implementace GBR z balíčku xgboost.

3.3.7 Ensembling

Ensembling (Hastie a kol., 2009) je technika, která kombinuje předpovědi více modelů s cílem zlepšit celkový výkon v regresní úloze. Ensembling může pomoci snížit přeučení, zlepšit zobecnění a zvýšit robustnost předpovědí. Je užitečné zejména tehdy, když se jednotlivé modely vzájemně doplňují a vyruší své chyby.Pro ensembling jsme zvolili průměr všech predikcí předchozích modelů.

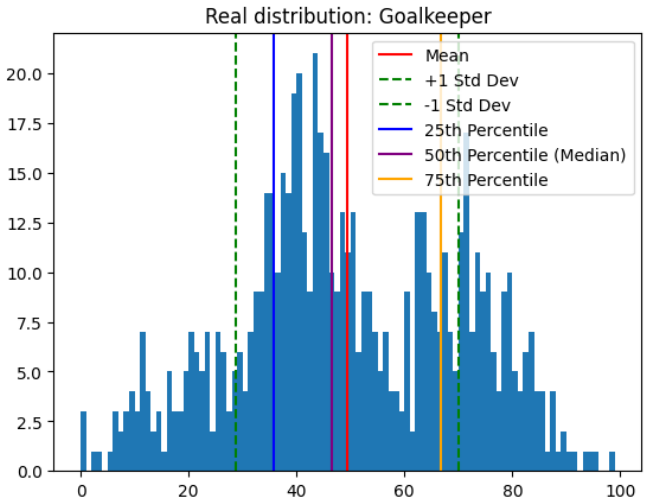
3.4 Ladění hyperparametrů

Použili jsme framework pro optimalizaci hyperparametrů, který se jmenuje Optuna (`https://optuna.org/`). Optuna provádí proces, kterému se říká trial, ve kterém vybere kombinaci hyperparametrů z rozsahů, které uživatel zadá k prozkoumání. Pro tuto kombinaci natrénuje model a vyhodnotí na validační množině pomocí zadané metriky (MSE). V tomto prostoru provádí hledání za pomoci heuristik a prořezávání a aproximuje tak nejlepší kombinaci. K prohledávání jsme zvolili prostory hyperparametrů, definované v příloze A.2.

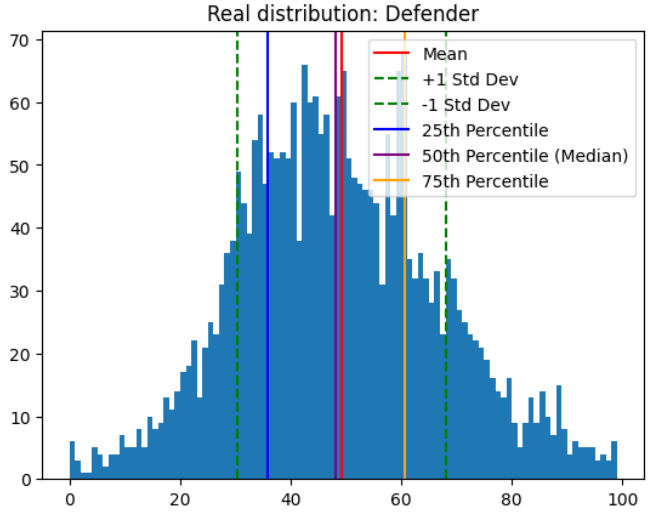
3.5 Rozdělení dat a rozbor cílové veličiny

Využíváme rozdělení v poměru 80 % trénovacích dat, 10 % validačních a 10 % testovacích. Na níže přiložených grafech 3.1, kde na ose x je *totalScore* a na ose y počet hráčů v testovací množině, lze pozorovat distribuce skóre pro jednotlivé hráčské pozice a vyvodit z nich informace užitečné pro hraní Sorare i pro potřeby práce.

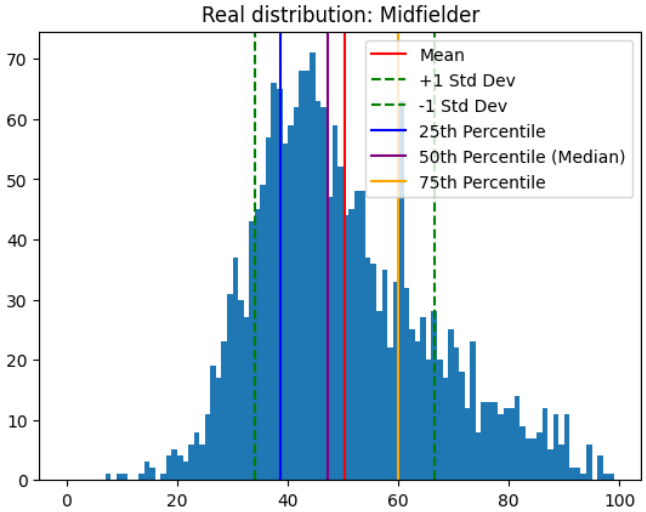
Nejnormálněji rozdělené je skóre pro obránce z důvodu, že ne tak často získávají *decisiveScore* (asistence, góly) a spíše u nich záleží na *allAroundScore*. Skóre útočníků i bránkářů je náchylné na změnu v *decisiveScore* z důvodu gólů, asistencí a u brankářů čistých kont a proto můžeme pozorovat bimodální rozdělení okolo dvou úrovní 35 a 60. U brankářů dochází k většímu zašumění a extrémnějším hodnotám (zákroky, chyby, 3 a více inkasovaných gólů viz. scoring matrix 1.2). U záložníků můžeme pozorovat normální rozdělení, ale oproti obráncům platí, že graf je v extrémních hodnotách nesymetrický ve prospěch pravé strany (je



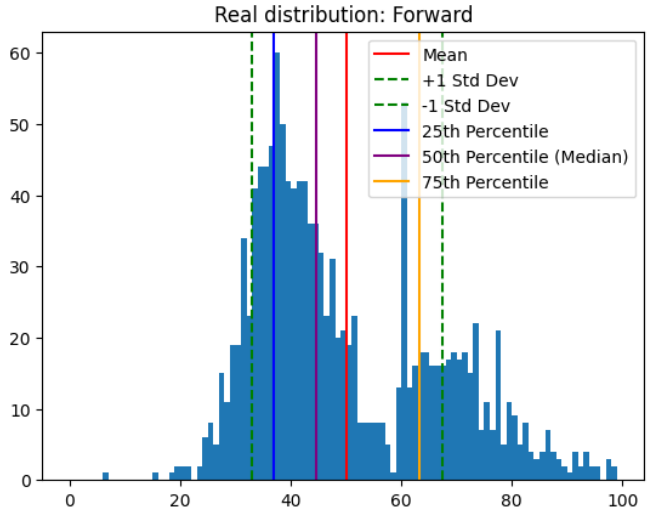
(a) Distribuce cílové veličiny pro brankáře



(b) Distribuce cílové veličiny pro obránce



(c) Distribuce cílové veličiny pro záložníky



(d) Distribuce cílové veličiny pro útočníky

Obrázek 3.1: Distribuce cílové veličiny pro jednotlivé hráčské pozice

běžnější získat 80 bodů, než 20) a unimodalita grafu poukazuje na větší důležitost *allAroundScore* oproti *decisiveScore*. 75. percentily v této sadě (definované tak, že 75 % skóre z této distribuce je menší než) jsou nejmenší u záložníků a obránců (60 oproti 65 u brankářů a 62 u útočníků) a poukazují na fakt, že tyto pozice obecně mívají menší skóre a tedy v případě, že získají vysoké skóre, dávají největší kompetetivní výhodu. Rozptyl je mezi pozicemi srovnatelný s výjimkou brankářů, kde je výrazně větší.

Prozatím jsme uvedli jako cílovou veličinu a učitele *totalScore*. To ale není podmínkou. Můžeme predikovat něco odvozeného, z čehož jde zpětně skóre dopočítat. Důležitou myšlenkou, která má potenciál zlepšit výsledky je to, že můžeme predikovat kolika násobek své L40 (průměr skóre z posledních 40 zápasů) hráč v daném zápase zahrál. Vycházíme z předpokladu, že když nevím nic o hráčově formě a o soupeři, tak nejlepším odhadem je L40 od které se skóre v jednotlivých odchyluje. Když poté přidáme kontextuální informace můžeme odhadnout jak se zápas bude pro hráče vyvíjet s tím, že známe jeho dlouhodobý průměr. Natrénovali jsem tedy modely, které předpovídají kolika násobek své L40 hráč v následujícím zápase zahraje na základě relevantních informací dostupných před zápasem, popsaných výše (nejdůležitější jsou informace o soupeři).

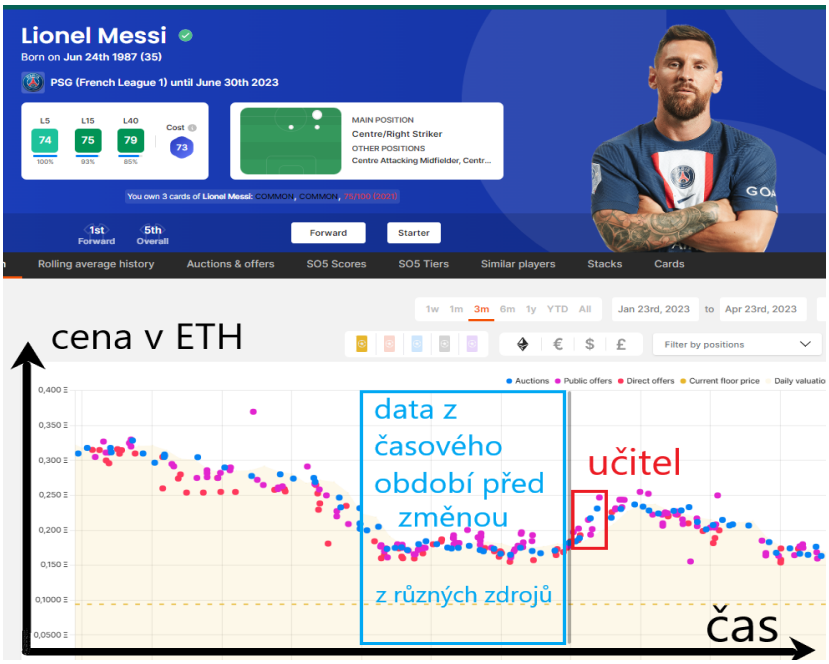
3.6 Sestavení týmu na základě predikcí

Když máme k dispozici odhady pro další výkony hráčů, je vhodné hráče, pro šanci dobrého umístění v žebříčku, využít tak, že maximalizujeme očekávaný počet bodů v rámci jedné sestavy. K tomuto je vhodný hladový algoritmus, který postupně z množiny dostupných hráčů vybere nejlepšího (s nejvyšším odhadovaným výkonem) brankáře, obránce, záložníka, útočníka a extra hráče. Tak sestaví „nejlepší“ tým. Opakuje pro druhý nejlepší, třetí nejlepší... Dokud jsou k dispozici hráči.

4. Určení vhodného nákupu karty

Problém určení vhodného nákupu karty je v zásadě identifikace vhodných hráčů. Každý hráč má sice karet až tolik, kolik je pro danou raritu možné, ale tyto karty si jsou ve své hodnotě v zásadě ekvivalentní. Rozdílem mezi nimi je rarita, sériové číslo, edice, což jsou faktory, které buď pro cenu důležité nejsou, nebo je pro tuto práci budeme fixovat. Pro fixaci byly zvoleny rare karty (těch je maximálně 100 ročně) z normální edice (speciální edice jsou z datasetu vyjmuty). Dále používáme pojem cena hráče ve smyslu průměrné ceny rare karty pro daného hráče za nějaké relevantní časové období, nebo za posledních n prodejů. Pokud identifikujeme podceněného hráče, je už poté lhostejné, kterou z jeho karet zvolíme ke koupi. Potom se dá problém chápat jako nalezení vhodné selekce f, pro zobrazení $f : X \to B$, kde X je množina všech hráčů a B je podmnožina hráčů vhodných ke koupi. K řešení jsme zvažovali dva různé přístupy:

- a) Mohli bychom predikovat další cenový vývoj hráče relativně vůči sobě samému (procentuální nárůst ceny v dalším měsící) a k nákupu označit ty hráče, pro které model predikuje dostatečně velký nárůst. Jako učitel k trénování by se použil procentuální nárůst/pokles ceny. Pro tento přístup by velkou důležitost měla data, zachycující změnu v relevantních parametrech. Změna počtu subscriptions na Sorare, změna sledujících na sociálních sítích, změna aktivity na platformě celkově, změna ve vyhledávanosti hráče na google, změna hráče v produktivitě (xG, xA, xGChain z Understat dat, nebo goalScored z league_stats), změna v hodnocení ve FIFA, změna v průměrném Sorare skóre a další. Z důvodu nedostatku dat strukturovaných pro sledování změn jsme tento přístup nezvolili, ale do budoucna by stál za vyzkoušení. Bylo by nutné data sbírat průběžně a ukládat si změny v daný časový moment.



Obrázek 4.1: Myšlenka přístupu a)

- b) Vytvořit množiny vzájemně si podobných hráčů a z těchto množin vybrat hráče, kteří jsou levnější než ostatní. Tento přístup využíváme a popisujeme v dalších sekcích.

4.1 Shlukování v přítomnosti

Využíváme kmeans (Hastie a kol., 2009) k vyrobení shluků hráčů na základě jejich objektivních kvalit v čase stažení dat. Příznaky pro shlukování budeme vytvářet ručně na základě toho, co je pro nás při nakupování karet důležité. V rámci výsledných shluků si spočítáme ceny hráčů a identifikujeme ty, kteří jsou podceněni oproti ostatním. Z důvodu nejasnosti, jak výsledky tohoto postupu chápat a vyhodnocovat, byl navržen experiment, který toto provádí v minulosti a má za cíl nálezt takové nastavení všech parametrů a výběr příznaků, které vede k dobrým výsledkům.

4.1.1 Využitý model: kmeans

Na připravených datech využíváme algoritmus kmeans a jeho implementaci z python balíčku sklearn. Kmeans je algoritmus strojového učení bez učitele, který rozřazuje vstupní data do předem určeného počtu shluků k na základě jejich vzdálenosti v prostoru příznaků. Algoritmus je iterativní a s datovými body pracuje tak, že je přiřadí do shluku, jehož centrum je nejblíž. Následně centra přepočítá na základě příslušných datových bodů. Proces se opakuje dokud nenastane maximální počet opakování, nebo konvergence. Kmeans je náchylný na nalezení suboptimálních řešení, proto k inicializaci využíváme kmeans++.

Protože jde o učení bez učitele, je složitější vyhodnocovat, jak kvalitně shlukování proběhlo. Inerce je ukazatel kvality, který reflektuje, jak jsou datové body blízko svým centrům shluků a nižší inerce je tedy lepší, protože máme kompaktnější shluky. Jelikož nemáme předem stanoveno k, využíváme elbow metodu. Elbow metoda je vykreslení grafu pro sledování závislosti inerce na počtu shluků a určení takového k, od kterého už inerce přestává prudce klesat a vyrovnává se.

4.1.2 Vytváření příznaků

Využíváme stejné zdroje jako pro regresi hráčských výkonů, ale v tomto případě je potřeba zachytit jak momentální formu hráče, kvalitu v rámci Sorare, tak i jeho obecnou kvalitu, popularitu a ohodnocení v reálném světě. Z game_stats využíváme průměrné *decisiveScore*, *allAroundScore* a odehrané minuty za posledních 10 a 30 zápasů. Tyto statistiky popisují kvalitu hráče v rámci Sorare, kde nejdůležitější je skóre a odehrané minuty - jestli hráč pravidelně nastupuje do zápasů a manažer s ním při umístění do sestavy může počítat. Z FIFA využíváme hodnotu hráče v eurech na reálném fotbalovém trhu a jeho *overall* (celkové) hodnocení, jeho *potential* hodnocení a také rozdíl mezi potenciálem a celkovým hodnocením pro rozeznání hráčů, kteří už svého potenciálu dosáhli a kteří mají největší fotbalový růst před sebou (podle FIFA). Z understat využíváme všechny statistiky popisující daného hráče (zejména očekávaná produktivita, produktivita, nególová produktivita). Z league_stats nevyužíváme žádná data, protože popisují hlavně tým daného hráče, což je faktor, který má určitě vliv na jeho

výkony, ale tato informace už je obsažena přímo v průměrech skóre nebo v tom, jaké jsou hráčovy individuální statistiky (když tým má menší kurz na výhru, tak si pravděpodobně vytváří více šancí, což zlepšuje hráčovo xG). Za prozkoumání stojí i závislost dalšího cenového vývoje na námi predikovaném příštím výkonu (výsledek kapitoly 3), ale toto v práci nezkoumáme.

Vytvořili jsme skupiny příznaků, kde každá skupina popisuje jiný aspekt kvality hráče. Konkrétní skupiny lze nahlédnout v tabulce 4.1.
Lr_ratio (limited to rare ratio) je příznak z ekonomické skupiny, který je roven podílu průměrné ceny limited/rare karet pro daného hráče. Trh s limited kartami je z důvodu nižších cen plynulejší a reaguje rychleji na změny a vývoj ve fotbalovém světě. Domníváme se, že abnormální poměr mezi těmito dvěma cenovými průměry by mohl signalizovat špatné nacenění a potenciální příležitost.
Down_from_ath (down from all time high) je příznak z ekonomické skupiny, který je roven podílu aktuální průměrné ceny ku nejvyšší zaznamenané ceně hráče (v rámci Sorare).

No.	Name	List of Features
1	All Around Score	allAroundScore*
2	Decisive Score	decisiveScore*
3	Krátkodobá forma	*L10*
4	Dlouhodobá forma	*L30*
5	Odehrané minuty	*mins_played*
6	FIFA kvalita hráče	overall, potential_diff
7	Ekonomika	value_eur, down_from_ath, lr_ratio
8	Očekávaná produktivita	xG, xA
9	Reálná produktivita	goals, assists
10	Ne-gólová produktivita	key_passes, xGChain, xGBuildup

Tabulka 4.1: Skupiny příznaků pro shlukování

4.1.3 Shrnutí postupu, problémy

Výsledný postup (selekce f) je tedy takový, že si na základě aktuálních dat vyrobíme datovou sadu, spustíme kmeans, které hráčům přiřadí shluk. Poté pro každý shluk vybereme hráče následujícím předpisem:

$$f(x) = \begin{cases} 1, & \text{pokud hráčova cena} < 0.5 * \text{median shluku} \\ 0, & \text{jinak} \end{cases} \tag{4.1}$$

Tímto způsobem můžeme vybrat hráče, kteří jsou levnější než zbytek jejich shluku a mohou být považováni za podceněné. Tento přístup má několik potenciálních problémů, jako je nemožnost okamžitého vyhodnocení. Bylo by nutné vyhodnocovat externě a na vyhodnocení čekat než se hráčova karta několikrát prodá pro získání cenového vývoje, což jsme ve vyhodnocení sice provedli, ale jen pro ilustrační účel. Dalším problémem může být, že ručně vybrané příznaky nejsou ideální pro identifikaci podceněných hráčů a třeba by lépe fungovala nějaká jejich podmnožina. Také je v této úloze předem neznámý počet shluků k,

pro jehož určení dobře nezafungovala elbow metoda. Výše zmíněné podtrhuje důležitost provádět další experimenty a validaci tohoto přístupu, aby bylo možné ověřit jeho účinnost a vyladit parametry a výběr příznaků pro dosažení lepších výsledků. Za tímto účelem jsme navrhli experiment popsaný v další sekci.

4.2 Shlukování v minulosti: Experiment

Pro shlukování v minulosti byly zvoleny časové úseky po dvou měsících od října 2021 do prosince 2022. Vyzkoušeli jsme všechny možné kombinace skupin z 4.1. Pro nastavení počtu shluků využíváme množinu K, která obsahuje jednotlivá k k vyzkoušení jako parametr kmeans. Množina byla nastavena jako

$$K = \{5, 10, 15, \ldots, \lfloor n/4 \rfloor\},$$

kde n je počet unikátních hráčů (počet dat).

Pro každé datum z časového úseku byla vyzkoušena každá kombinace příznaků a vyzkoušena všechna k z K, kde jsme zvolili to vhodné pomocí elbow metody, která pro pro menší počet příznaků fungovala spolehlivě. Pokud pomocí ní vhodné k nalezeno nebylo, bylo zvoleno to s nejnižší inercí. Pro všechny kombinace popsané výše byl natrénován kmeans model, který hráče rozřadil do shluků. Tomuto jednomu procesu říkáme běh. Pro každý běh byli identifikováni podcenění hráči (opět jako levnější než polovina mediánu jejich shluků). Pro tyto hráče jsme dále spočítali, jak se jejich cena vyvíjela v dalším měsíci a vyhodnotili relativně vůči svému shluku v sekci 6.2.

5. Vyhodnocení predikce výkonů

5.1 Optikou RMSE

Jako hlavní metriku využíváme RMSE, která je definovaná jako odmocnina z průměrného čtverce rozdílu predikce od reálné testovací hodnoty. Matematicky se dá vyjádřit takto:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{5.1}$$

kde n je počet testovacích dat, y_i je reálná hodnota a $\hat{y}_i$ je predikovaná hodnota. RMSE je oblíbená metrika pro vyhodnocování regresních modelů z důvodu dobré interpretability, jelikož nám jasně říká, jak moc se model průměrně ve svých predikcích mýlí.

Nejlepší dosažené RMSE byly 16,41; 15,43; 17,8; 20,3 pro útočníky; záložníky; obránce; brankáře, což je ve všech případech lepší, než dosáhl model predikující průměr, ale horší než byla očekávání. I přes velký počet natrénovaných modelů a vyzkoušených postupů se od určitého bodu RMSE jen těžko snižovala.

Ve zbytku sekce sledujeme, jaký vliv měly využité metody na výsledky a k tomu využíváme dva typy grafů. V prvních grafech platí, že modely jsou seřazeny vzestupně podle rostoucí RMSE. Svislá přerušovaná čára značí, kde se na dané škále umístil model predikující vždy průměr pro danou pozici. Grafy jsou rozděleny po pozicích a pro každou pozici jsou sloupce stejné, jen se na výsledky díváme různými optikami pro zhodnocení, které techniky si vedly nejlépe. Tyto optiky jsou: použitý model, predikce odchylky od L40 oproti predikce přímo skóre, nastavení/nenastavení hyperparametrů, použitý výběr příznaků. Pro další porovnání použitých metod využíváme i druhý graf, kde na ose y je průměrná chyba přes všechny modely, pro které byla použita metoda na ose x.

1. Modely
   Obecně pro všechny pozice platí, že nejlépe fungovala LR (hlavně u útočníků a brankářů je dominujícím modelem v 10% nejlepších viz. grafy 5.1d, 5.1a) následovaná MLP u brankářů, SVR u obránců (graf 5.1c), ensemble a RF u záložníků (graf 5.1b) a MLP a ensemble u útočníků. U brankářů bylo nejtěžší porazit průměrnou predikci kvůli velké variabilitě v distribuci cílové veličiny (obrázek 3.1a), nadprůměrných modelů bylo výrazně menší procento než u jiných pozic. Na každém grafu si lze všimnout, že MLP zabírají pravou stranu, tedy že často mají velkou RMSE. To může být způsobeno nedostatečným důrazem na návrh architektury neuronové sítě, kde byly prozkoumány jen 4 varianty o maximálně 4 vrstvách.

2. Feature selection
   Pro všechny pozice a pro záložníky nejvíce platí (viz. graf 5.3), že nejlépe funguje selekce na bázi random forest. U brankářů se ukázala jako vhodná i selekce na základě korelací (viz. graf 5.2). PCA je nejméně vhodná, protože nikdy nevedla k nejlépe predikujícím modelům, ale zase neplatí, že by

byla dominantní u modelů nejhorších. Na grafu 5.4 můžeme pozorovat seřazení rf, corr, pca, a nepoužité selekce příznaků v rámci průměrné chyby pro útočníky. Podobné seřazení, byť ne tak lineární, platí i pro ostatní pozice. Pro defenzivní pozice se metoda corr skoro vyrovnala metodě rf (viz. graf 5.5). Nikde v rámci výběru příznaků neřešíme vzájemnou korelaci mezi nimi. Za tímto účelem jsme využili ruční selekci, která si brala inspiraci z výsledků metod rf a corr, ale zároveň se snažila odebrat korelované příznaky. U tří pozic (útočník, obránce, brankář) se tak povedlo dosáhnout ještě lepších výsledků **(16,36; 17,79; 20,06)** . Jaké příznaky byly využity a jejich vysvětlení uvádíme v příloze A.3.

3. Cílová veličina
   Pro brankáře je na grafu 5.6 celkem jednoznačně vidět, že většina nadprůměrných modelů predikovala rovnou *totalScore*. Ukazujeme také graf 5.7 pro záložníky, kde je levá strana grafu od průměru nejčervenější oproti ostatním pozicím. Pozice záložníka tedy oproti jiným více benefituje z predikování odchylky od L40. Ačkoliv grafy pro většinu ostatních pozic jsou neprůkazné, v průměrném případě se ukázalo, že je výhodnější predikovat standardně, tedy *totalScore* a ne odchylku od hráčovy L40, to platí i pro záložníky 5.8. Toto zjištění je překvapující, protože původně panovala hypotéza, že predikce odchylky dosáhne lepších výsledků.

   Na druhou stranu to říká, že velký vliv mají jiné faktory než hráčova obecná kvalita jako například proti jakému soupeři hraje a ostatní okolnosti zápasu. Hráčova L40 je nejviditelnějším faktorem, který manažeři na Sorare zkoumají a důležitost jiných faktorů podtrhuje možnost využití dat a strojového učení pro dosažení lepších odhadů.

4. Ladění hyperparametrů
   Ladění hyperparametrů se neukázalo jako důležité, kde průměrné chyby pro modely s/bez nastavení hyperparametrů si byly velice podobné v řádu tisícin, setin a pro brankáře desetin. To může být z důvodu nedostatečného prohledávaného prostoru hyperparametrů (hlavně v případě MLP).

Závěrem uvádíme pro každou pozici několik dobře predikujích modelů a jejich nastavení. Každý řádek, který se v tabulce 5.1 objeví, je to nejlepší nalezené nastavení pro daný model. V této tabulce značíme sloupec pro feature selection jako FS a sloupec pro parametry threshhold a n (popsané v sekci 3.2) jako th/n.

Tabulka 5.1: Nejlepší modely pro každou pozici a jejich nastavení

Pozice	**RMSE**	**Model**	**FS**	**th/n**	**Hyper**	**Target**
Forward	16.41	RidgeCV	rf	0.15	ne	totalScore
	16.52	MLP	rf	5	ne	totalScore
	16.53	ensemble	rf	50	ne	devFromL40
	16.53	RF	rf	0.15	ne	totalScore
	16.65	SVR	rf	25	ano	devFromL40
	16.79	XGBR	rf	5	ano	devFromL40
Midfielder	15.43	RidgeCV	rf	0.15	ne	totalScore
	15.52	ensemble	rf	50	ano	devFromL40
	15.53	RF	rf	0.15	ne	devFromL40
	15.60	MLP	rf	40	ne	devFromL40
	15.63	SVR	rf	0.1	ne	totalScore
	15.75	XGBR	rf	50	ano	devFromL40
Defender	17.80	RidgeCV	corr	50	ne	totalScore
	17.84	SVR	rf	0.1	ne	totalScore
	17.85	RF	rf	0.1	ne	totalScore
	17.94	MLP	rf	15	ne	devFromL40
	17.95	ensemble	rf	40	ne	devFromL40
	18.29	XGBR	corr	0.1	ano	devFromL40
Goalkeeper	20.29	RidgeCV	corr	0.1	ne	totalScore
	20.33	MLP	rf	30	ne	totalScore
	20.40	RF	rf	50	ne	totalScore
	20.44	ensemble	rf	50	ne	totalScore
	20.48	SVR	rf	5	ano	totalScore
	21.21	XGBR	corr	0.2	ne	totalScore

Pro modely, které vybíraly nižší počet příznaků poskytujeme jejich výčet a komentář, ostatní lze nalézt v elektronické příloze práce. Pro útočníky se ukázalo jako 5 nejdůležitějších příznaků

```
LS_L40_goalsScored_all_mean,
L40,
gs_player_mean_successful_final_third_passes,
gs_player_mean_allAroundScore,
_LS_L15_goalsConceded_all_mean
```

což jsou příznaky popisující dlouhodobu kvalitu a kvalitní snažení na útočné třetině. Také je přirozeně důležité, kolik gólů dostává soupeřův tým v průměru.

Pro obránce máme v nejlepších modelech neuronovou síť s 15 příznaky

```
LS_L15_avgWinOdds_all_mean, LS_L15_avgWinOdds_home_mean,
L40, is_home, enemy_LS_L15_avgWinOdds_all_mean,
enemy_LS_L15_maxWinOdds_all_mean,
enemy_LS_L15_avgWinOdds_home_mean,
enemy_LS_L15_avgWinOdds_away_mean,
enemy_LS_L15_goalsScored_all_mean,
gs_player_mean_allAroundScore, gs_player_mean_accurate_pass,
```

```
gs_player_mean_poss_won, gs_player_std_poss_won,
gs_player_mean_interception_won, gs_player_std_interception_won,
```

Je vidět, že se tato síť snažila modelovat domácí/hostující prostředí, neboť velká část příznaků obsahuje slova "home"a "away"a zároveň je přítomen indikátor is_home. Za tímto účelem brala v potaz hlavně průměrné a maximální kurzy sázkových kanceláří na výhru ať už svého nebo soupeřova týmu. Průměrný počet získaných míčů a překažených přihrávek bychom pro obránce čekali jako důležitý faktor, ukázalo se ale, že důležitý je i rozptyl v těchto atributech. Ten může signalizovat schopnost hráče zahrát vysoké skóre a být aktivním činitelem v defenzivě v případě, že je tým pod náporem.

Pro brankáře byly vybrány tyto příznaky:

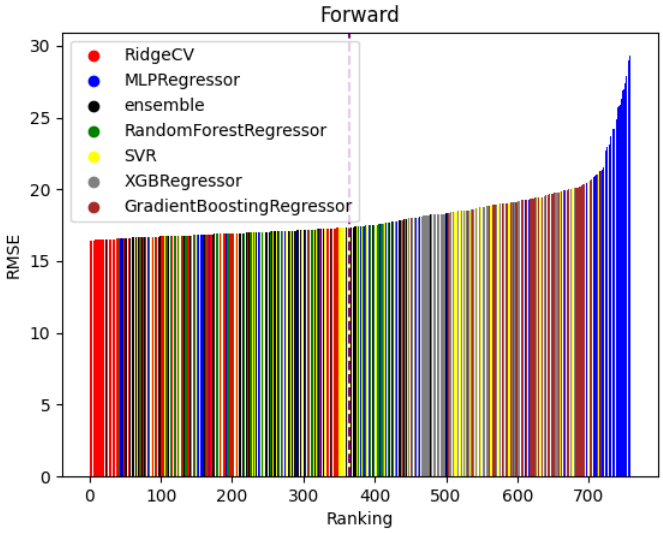
```
enemy_LS_L15_goalsScored_all_mean,
enemy_LS_L40_goalsScored_all_mean,
LS_L15_goalsConceded_all_mean,
LS_L40_goalsConceded_all_mean,
LS_L40_goalsConceded_away_mean
```

které porovnávají pouze vstřelené góly soupeřova týmu a obdržené góly hráčova týmu v různých časových úsecích.

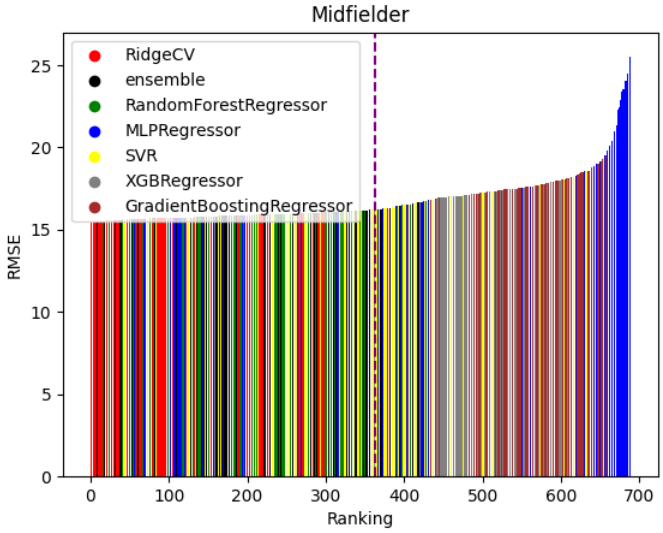
Pro záložníky se nejlépe umístil model s rf výběrem, kde příznaky mají přes 0,15 důležitosti a je jich velký počet. Uvádíme jen několik nejdůležitějších a těch, které jsou specifické pro záložníky:

```
u_teammates_xGChain,
u_player_assists,
gs_player_mean_successful_final_third_passes,
gs_player_mean_accurate_pass,
gs_player_mean_accurate_long_balls,
enemy_LS_L15_shotsAgainst_all_mean,
u_player_key_passes,
u_teammates_xGBuildup
```

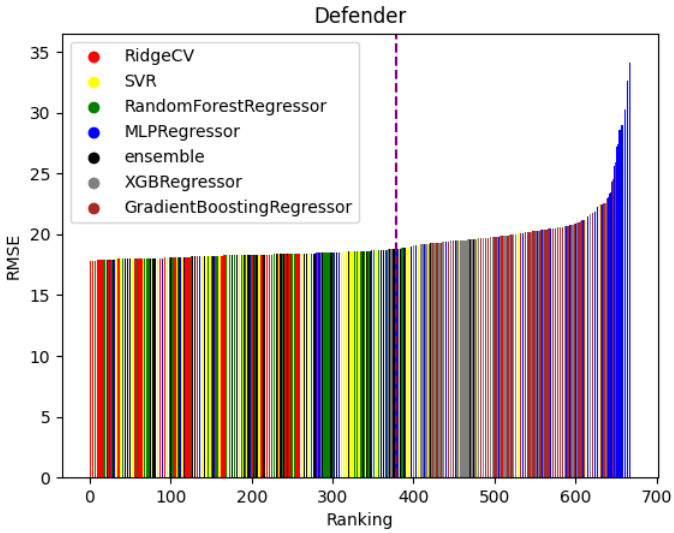
Zde panuje největší dominance příznaků z Understat, hlavně statistiky jako xGChain, xGBuildup (definované v sekci 2.3.4 o datech z Understat) všech spoluhráčů a také hráčovo kreativní snažení jako přihrávky, dlouhé přihrávky, klíčové přihrávky, asistence. Odlišným ukazatelem v nejlepších příznacích je průměrný počet střel, které na sebe soupeřův tým nechal vyslat v posledních 15 zápasech. Za zmínku stojí, že v příznacích na pozicích 15-30 v důležitosti se objevilo 10 spočítaných jako standardní odchylka z nějaké důležité statistiky (např. gs_player_std_pen_are_entries), což je proti očekáváním. Očekávali jsme přítomnost např. odchylky pro skóre, která když je větší, značí schopnost hráče zahrát vysoké skóre. Důležitost odchylky přes tak velké množství statistik je překvapivé.



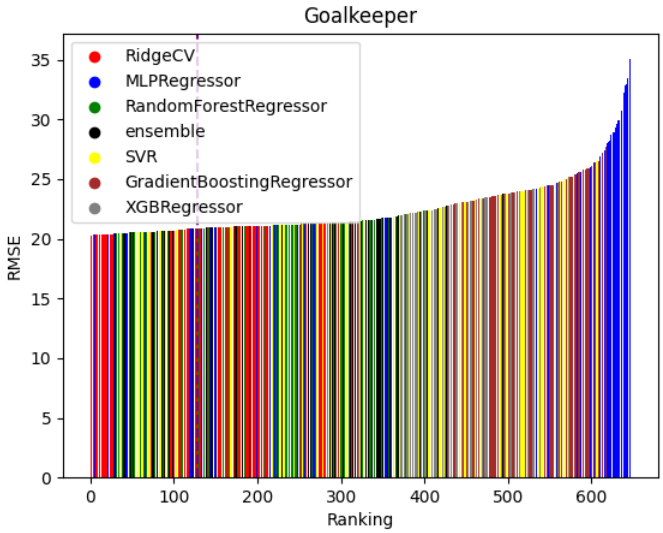
(a) Všechny modely pro pozici forward



(b) Všechny modely pro pozici midfielder

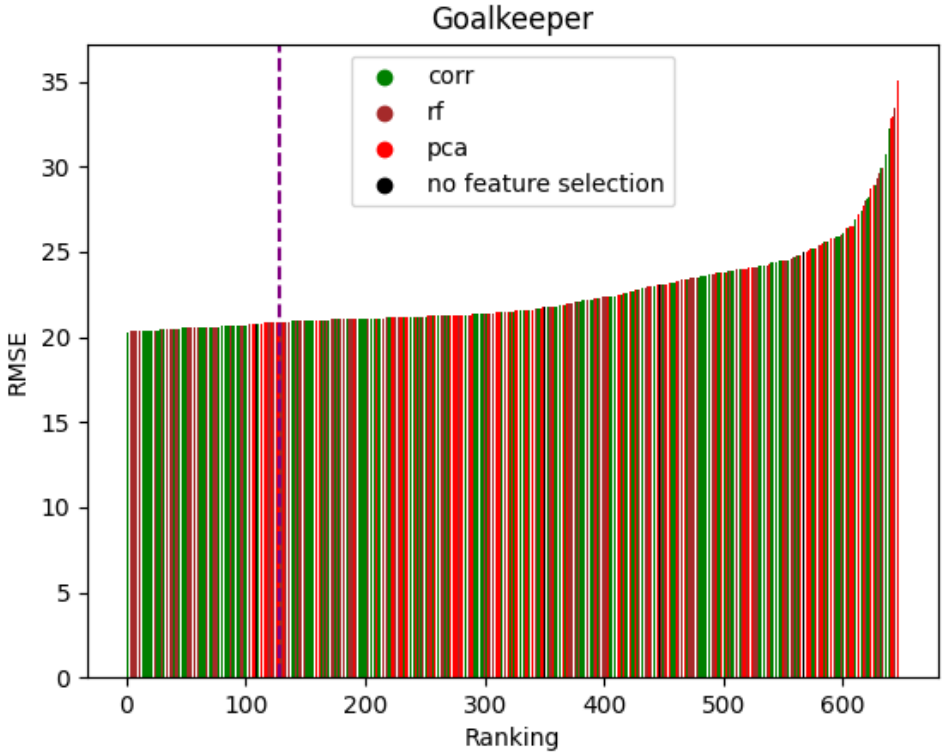


(c) Všechny modely pro pozici defender

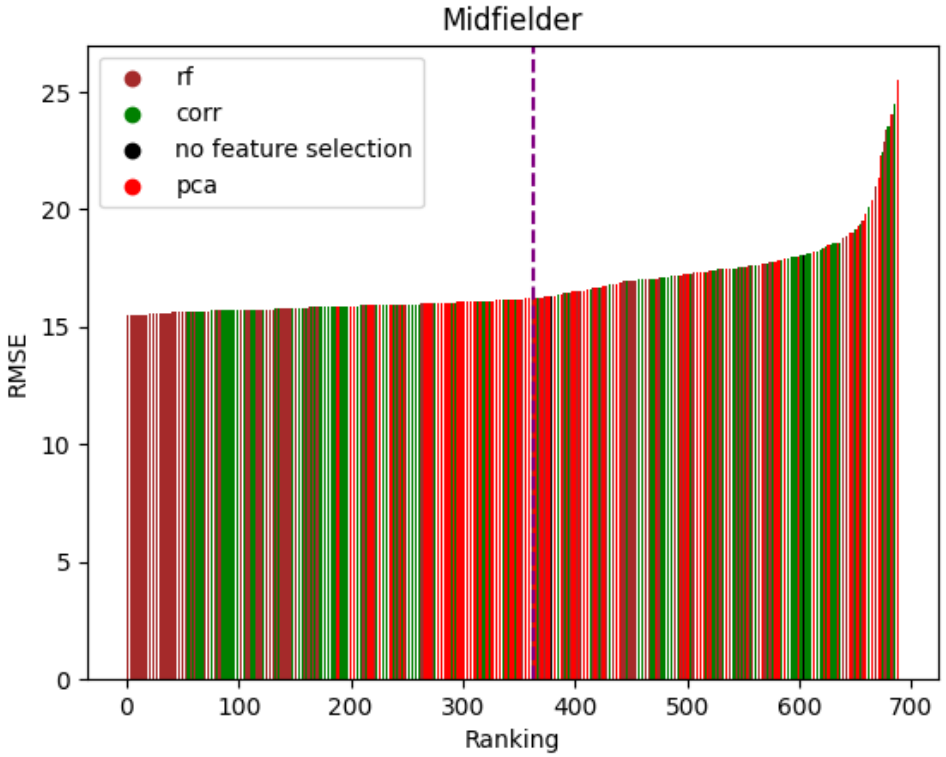


(d) Všechny modely pro pozici goalkeeper

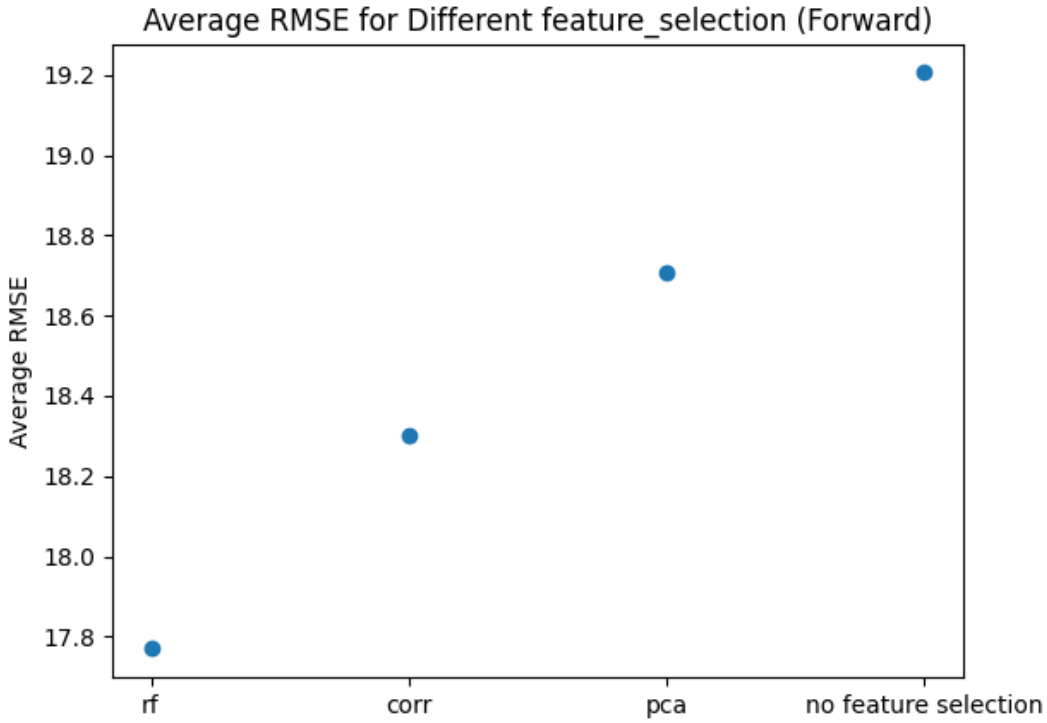
Obrázek 5.1: Všechny modely pro jednotlivé hráčské pozice



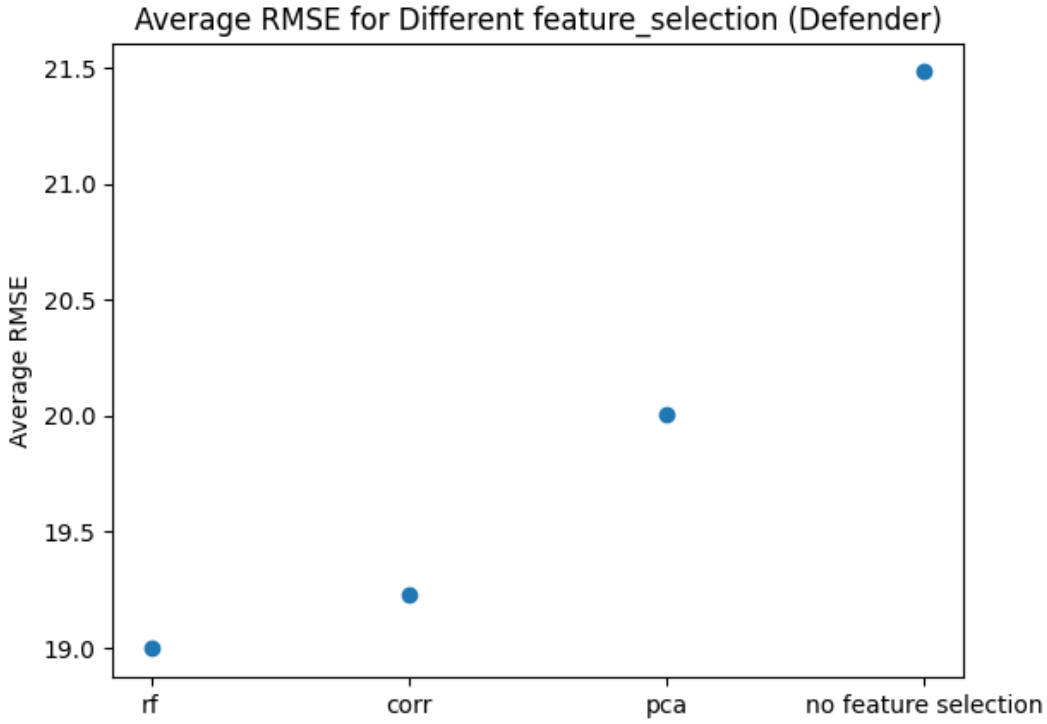
Obrázek 5.2: Všechny selekce pro pozici goalkeeper



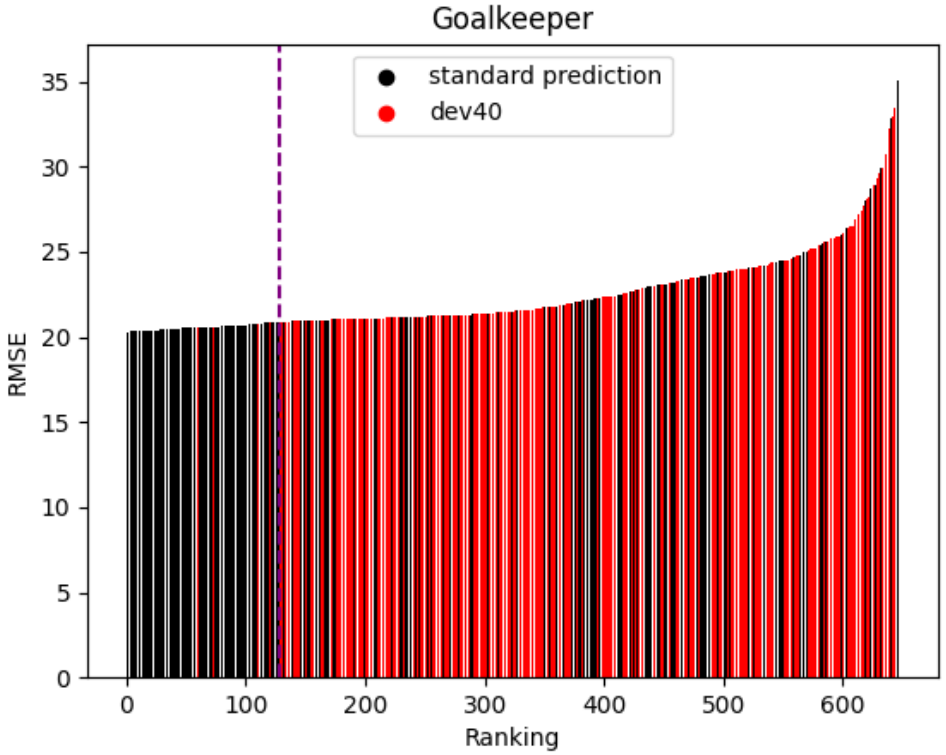
Obrázek 5.3: Všechny selekce pro pozici midfielder



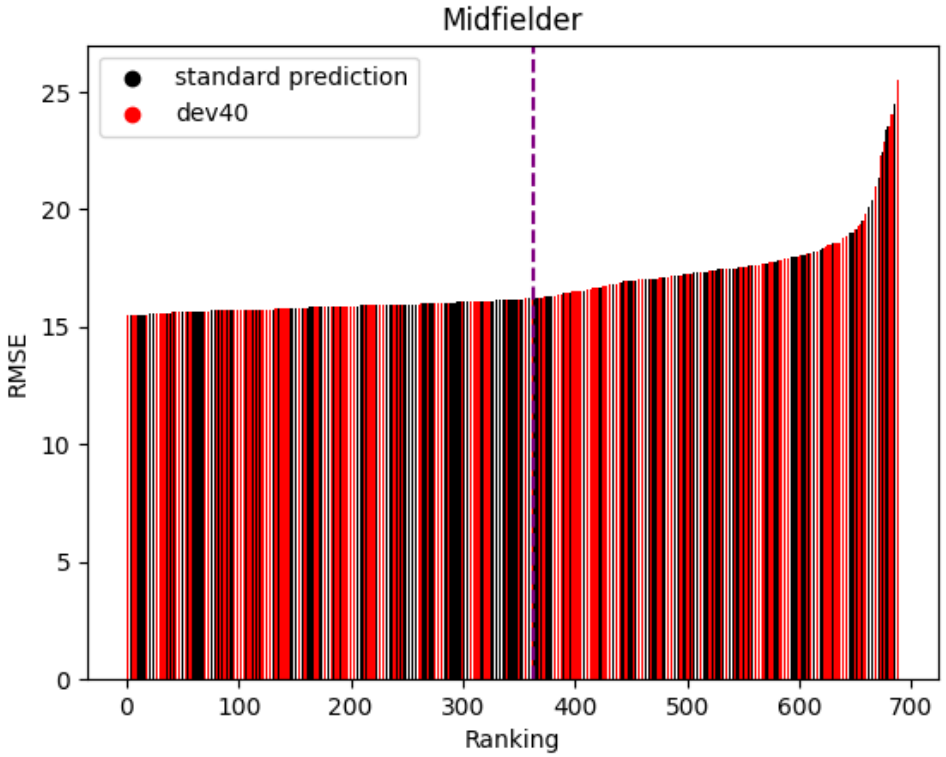
Obrázek 5.4: Průměry RMSE pro selekce pro pozici forward



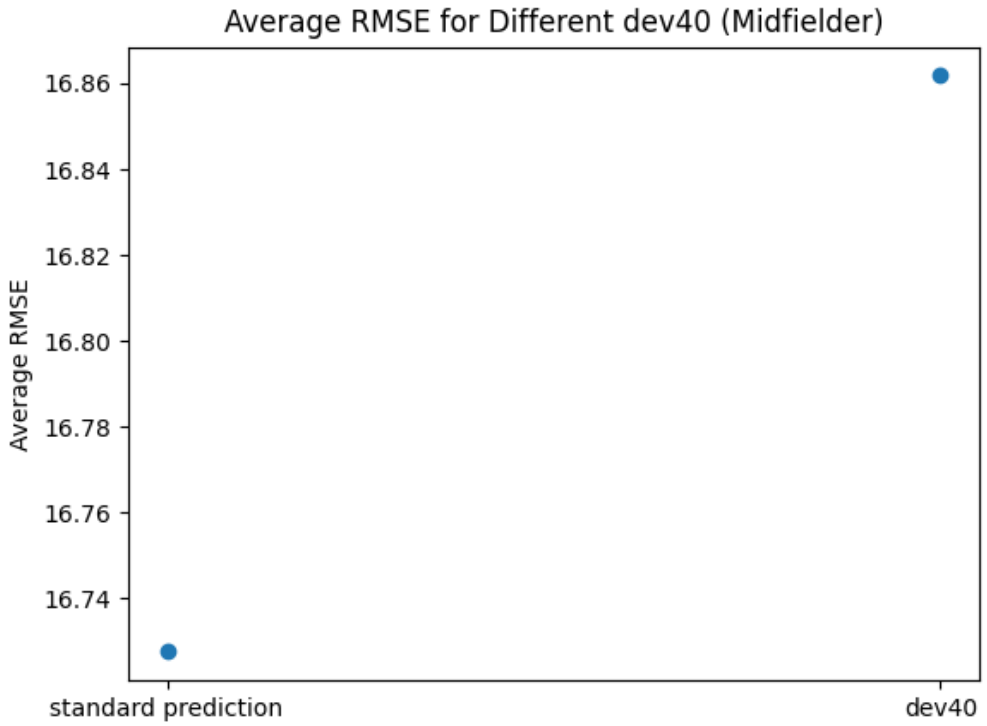
Obrázek 5.5: Průměry RMSE pro selekce pro pozici defender



Obrázek 5.6: Cílová veličina pro pozici goalkeeper



Obrázek 5.7: Cílová veličina pro pozici midfielder



Obrázek 5.8: Průměry RMSE pro cílové veličiny pro pozici midfielder

5.2 Optikou poznávání výjimečných výkonů

Využití RMSE je vhodné, pokud je naší motivací pro všechny hráče minimalizovat chybu predikce. Při hraní Sorare je ale důležitější modelovat, když hráč zahraje výjimečně dobře z důvodu velké kompetetivnosti soutěží. Pro dobré umístění a s tím spojené ceny je běžně potřeba získat 400 a více bodů, což dělá průměr 80 na jednoho hráče. Při využití standardních vah pro jednotlivé trénovací body se modely v rámci minimalizování chyby naučí predikovat okolo průměru a jen málokdy jsou schopny rozpoznat nadcházející dobrý výkon. Z toho důvodu nám tolik nevadí predikovat průměr, když hráč zahrál průměrně či podprůměrně jako v případě, kdy zahrál nadprůměrně.

Pro docílení větší smělosti modelů jsme zvolili metodu ručního nastavení vah pro trénovací body, sample weights. Matematicky se pak ztrátová/vyhodnocovací funkce dá vyjádřit jako

$$\text{Weighted RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}, \tag{5.2}$$

kde w_i je váha přiřazená i-tému datovému bodu. K nastavení vah jsme využili funkci vzdálenosti od průměru. Čím je bod více vzdálený, tím je pro trénink důležitější a funkce navíc přiřazuje větší váhy pozitivním příkladům. Dá se vyjádřit takto:

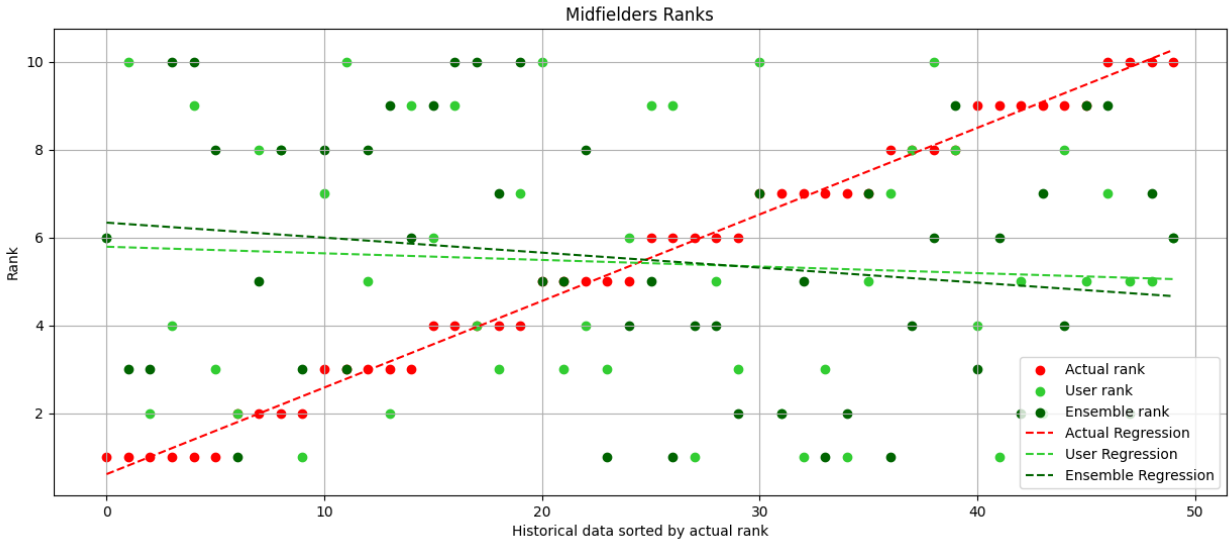
$$w_i = \begin{cases} \text{positive\_scaling\_factor} \cdot \frac{|y_i - \bar{y}|}{\max_j |y_j - \bar{y}|} & \text{pokud } y_i > \bar{y} \\ \frac{|y_i - \bar{y}|}{\max_j |y_j - \bar{y}|} & \text{jinak} \end{cases}, \tag{5.3}$$

kde w_i je sample weight pro i-tý bod, y_i je reálná hodnota, $\bar{y}$ je průměr cílových hodnot a positive_scaling_factor je konstanta využitá pro prioritizaci pozitivních výjimečných výkonu nastavená na 2.

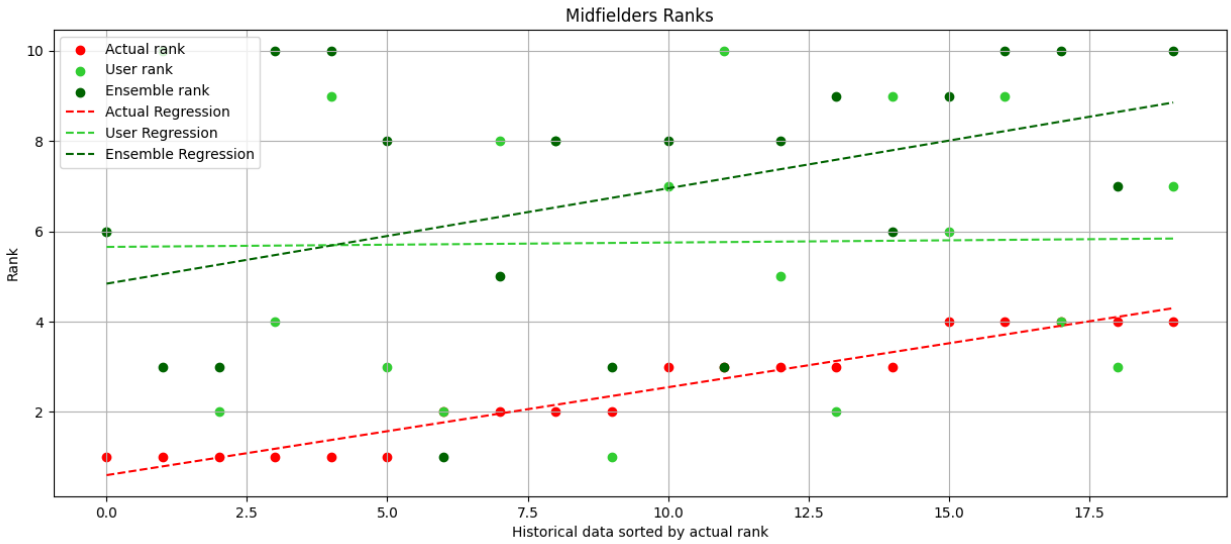
Využití této metriky je výhodnější pro reálné využití z výše popsaných důvodů, a proto byla v průběhu práce sbírána data pro zhruba 50 různých hráčů v 5 různých herních týdnech. Ke každému hráči byla poznamenána predikce autorů a následné reálné zahrané skóre. Cílem bylo na reálných datech zjistit, jestli jsou modely schopné detekovat dobré výkony a jestli by na základě toho byly sestaveny lepší týmy než sestavili autoři práce. K vyhodnocení používáme ensemble (tedy průměr přes všechny modely). Pro každou pozici bylo zvoleno nastavení toho ensemble modelu, který měl nejmenší čtvercovou chybu a s tímto nastavením byl model přetrénován s metrikou *weighted RMSE*.

Pro každou pozici přirozeně vzniká seřazení hráčů (rank) podle odhadovaného skóre a to jak pro reálné skóre, autorskou a ensemble predikci. Toto seřazení použijeme k vyhodnocení a ptáme se, jak je ensemble model schopný seřadit hráče oproti člověku. Níže si zobrazíme graf, kde každý bod na ose x je záznam v nasbíraných datech. Každému záznamu přísluší tři body: reálný rank, uživatelský rank a ensemble rank. Přerušované přímky jsou potom lineární regrese přes příslušné body. Data byla před zobrazením seřazena podle ranku, což lze pozorovat na červených bodech a křivce. Čím více se křivky blíží té červené (reálný rank), tím lépe člověk/model hráče seřadil.

V grafu 5.9 je vidět, že člověk seřadil hráče lépe než ensemble. Na začátku sekce jsme ale říkali, že důležité je identifikovat ty nejlepší hráče, které máme k dispozici. V tomto ohledu byl ensemble model dokonce lepší, což je vidět pokud graf (5.10) přiblížíme vyfiltrováním jen 20 nejlépe umístěných hráčů. K ilustraci byly použiti záložníci, ale pro ostatní pozice platí to samé.



Obrázek 5.9: Vyhodnocení weighted RMSE

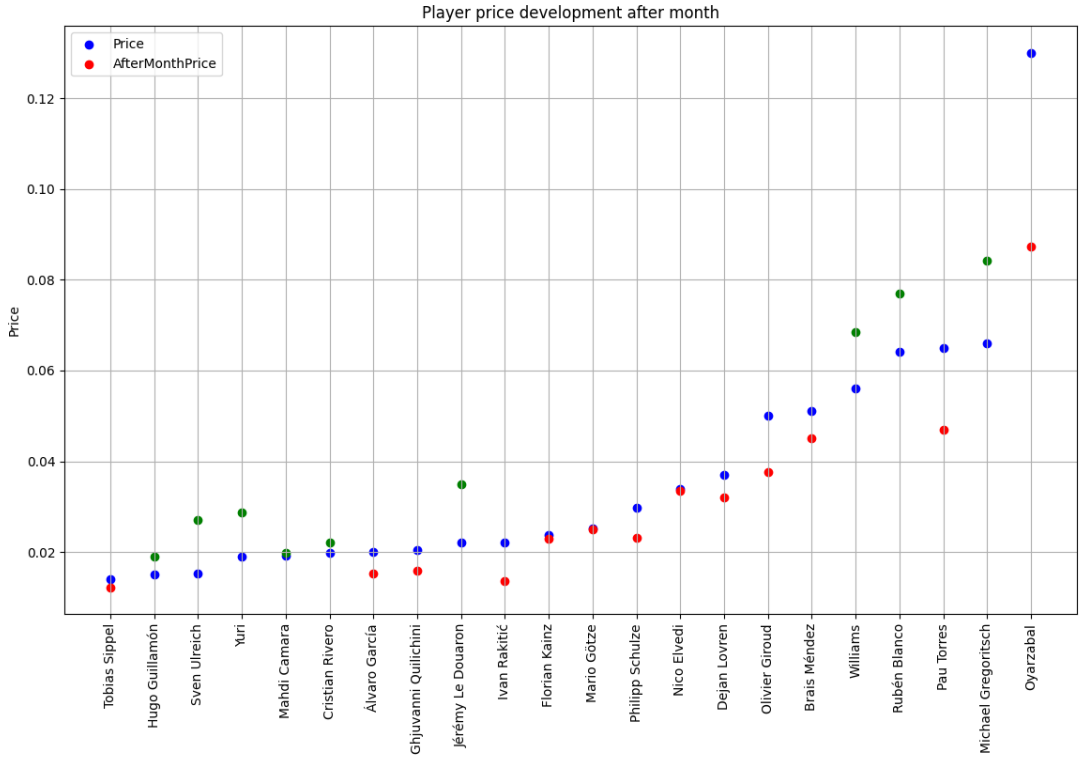


Obrázek 5.10: Vyhodnocení weighted RMSE pro výjimečné výkony

6. Vyhodnocení identifikace podceněných hráčů

6.1 Shlukování v přítomnosti

Začátkem dubna jsme spustili algoritmus popsaný v sekci 4.1 a zaznamenali hráče, které vybral. U těchto hráčů jsme spočítali, jaká byla jejich průměrná cena na začátku května, tedy o měsíc později. Výsledek lze pozorovat v grafu 6.1. Toto vyhodnocení podléhá vlivu trhu, protože probíhalo od dubna do května, kdy se evropské soutěže chýlí ke konci a typicky u všech karet klesá cena. V tomto období lze za úspěch považovat to, že karta zaznamenala jen mírný pokles, nebo si dokonce udržela svou cenu. To totiž znamená, že si na krátkém časovém úseku vede lépe než jiné karty. Tento relativní pohled, kdy bereme do úvahy vývoj ceny karty vůči jiným kartám je pohled, který je méně náchylný externím vlivům (cena ETH, počet manažerů na Sorare, sezónnost) a který dále prozkoumáváme v 6.2.



Obrázek 6.1: Externí vyhodnocení shlukování

6.2 Shlukování pro pohled do minulosti: experiment

Jedno shlukování je určeno pozicí hráče, datumem, kombinací zvolených skupin příznaků a hodnotou k. Pro všechny kombinace parametrů popsané výše byl natrénován kmeans model, který hráče rozřadil do shluků. Ve shlucích byli

identifikováni hráči s cenou pod polovinou mediánu daného shluku. Pro danou množinu hráčů jsme sledovali vývoj jejich ceny v dalším měsíci. Každý takovýto hráč zabíral v původním shluku nějaké procento celkové ceny a nějaké procento zabírá i z celkové ceny po měsíci. Zvolenou metrikou pro ohodnocení konkrétního shlukování je potom průměrný rozdíl nového procenta od původního přes všechny podceněné hráče. Popíšeme tuto metriku formálněji.
Nechť B je množina podceněných hráčů vybraných algoritmem. Každý hráč je součástí shluku. Pro každého hráče $i \in B$ značí x_i, kolik procent z celkové ceny shluku patřilo hráči i a $\hat{x}_i$ je to stejné, ale o měsíc později. Pak se metrika spočítá jako rozdíl mezi $\hat{x}_i$ and x_i pro všechny hráče z B. Metrika se dá matematicky vyjádřit takto:

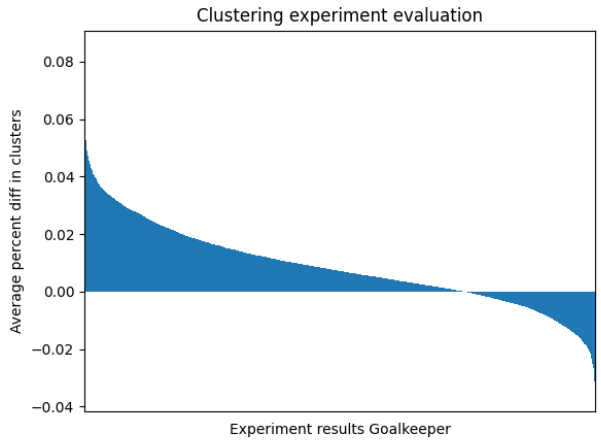
$$\text{Avg Percent Diff} = \frac{1}{|B|} \sum_{i \in B} (\hat{x}_i - x_i) \tag{6.1}$$

Nejlépe algoritmus fungoval pro **brankáře**, kde největší průměrný procentuální nárůst byl zhruba 8 % pro běh daný parametry 2022-12-31, skupiny [2, 4, 5, 6, 8, 9] a k=20. Pro **útočníky** byl nejlepší běh s parametry 2022-10-31, skupiny [2, 6, 7] a k=15 s nárůstem 3 %. Pro **obránce** byl nejlepší běh horší a to s nárůstem 2,5 %. Parametry byly nastaveny takto: 2022-08-31, skupiny [1, 3, 4, 6, 8, 10] a k=25. Ještě horší byl nejlepší běh pro **záložníky** s parametry 2022-10-31, skupiny [1,6,8,9] a k=30 s nárůstem pouze 1,3 %. Je nutné podotknout, že brankářů je zhruba 4x méně než obránců a 2,5x méně útočníků a přirozeně je pro ně průměr citlivější vůči extrémům. Z tohoto důvodu je důležitý pohled na distribuce výsledků pro jednotlivé pozice.

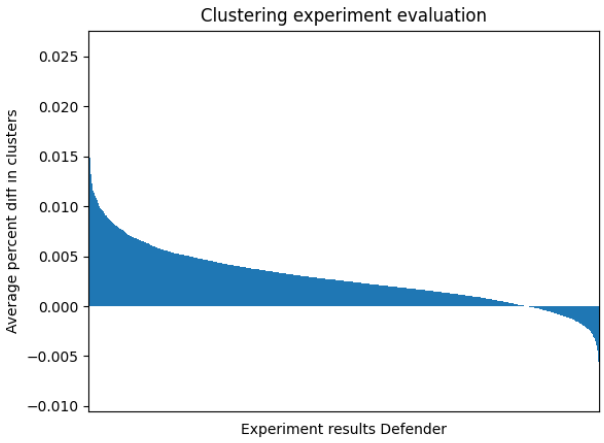
Pro jednotlivé pozice uvádíme níže grafy, ze kterých je vidět, že navrhnutý algoritmus dosahuje v rámci naší metriky pozitivních výsledků. V těchto grafech je na ose y uvedená metrika v desetinném čísle (0.01 odpovídá 1 %) a na ose x jednotlivé běhy shlukování. Pro všechny pozice platí, že nejlepší dosažený výsledek je v absolutní hodnotě větší než ten nejhorší a distribuce je výrazně nakloněná na pozitivní stranu.

Při zkoumání nejlepších běhů a jejich parametrů vyvstávají otázky jako jak velké k vedlo k nejlepším výsledkům, jaké skupiny příznaků vedly k nejlepším výsledkům apod. Na tyto otázky odpovídáme spočítaním sumy procentuálních nárůstů všech běhů, které v nastavení měly daný parametr. Pro parametry, které nejsou ve výsledcích pravidelně a deterministicky zastoupeny jako je k a pozice, uvádíme průměr. Ze skupin se ukázala jako ta nejklíčovější ekonomická. Ta obsahuje reálnou hodnotu hráče, *lr_ratio* a také *down_from_ath* (oba příznaky popsané v podsekci 4.1.2). Poslední jmenovaný příznak může poukazovat na důležitost toho, že při nákupu karty za větší cenu, než byl aktuální průměr, slouží psychologický efekt toho, že už ji někdo koupil ještě dráž, jako ospravedlnění tohoto rozhodnutí. Další důležitou skupinou byla kvalita hráče ve FIFA, což je objektivní a dlouhodobé hodnocení hráčových kvalit. Tyto dvě informace mohou naznačovat, že dobrou strategií je čekat na pokles cen známých a kvalitních hráčů a při tomto poklesu nakupovat.

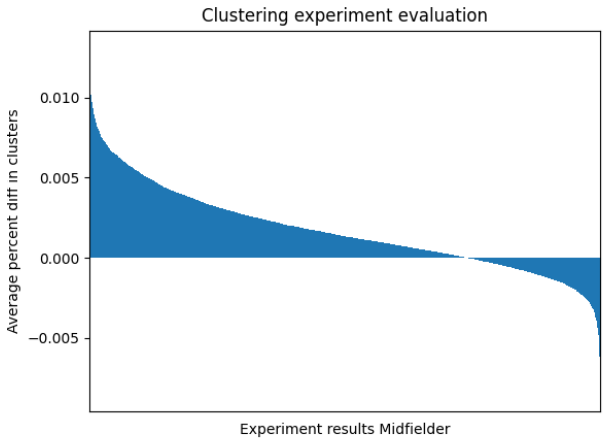
Co se týče pozic, tak dvě z nich, které dosáhli největších procentuálních nárůstů ve vyhodnocení, na tom nejsou průměrně až tak dobře (útočník, obránce). To je v souladu s faktem o počtu hráčů pro každou pozici, který jsme uvedli výše Na nejlepší průměrnou procentuální změnu dosáhli obránci, kterých je v datové



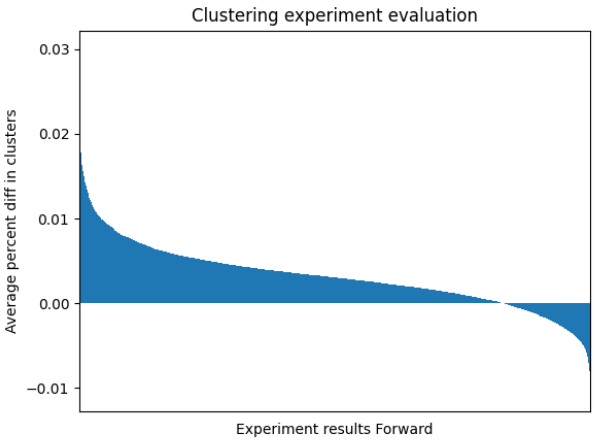
(a) Vyhodnocení shlukování pro brankáře



(b) Vyhodnocení shlukování pro obránce



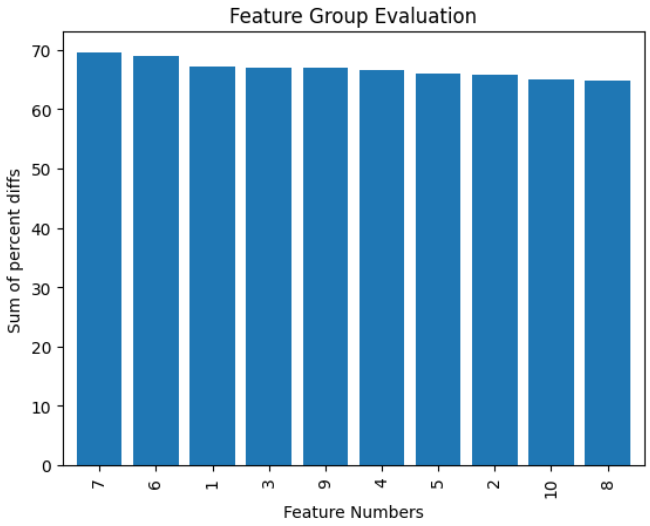
(c) Vyhodnocení shlukování pro záložníky



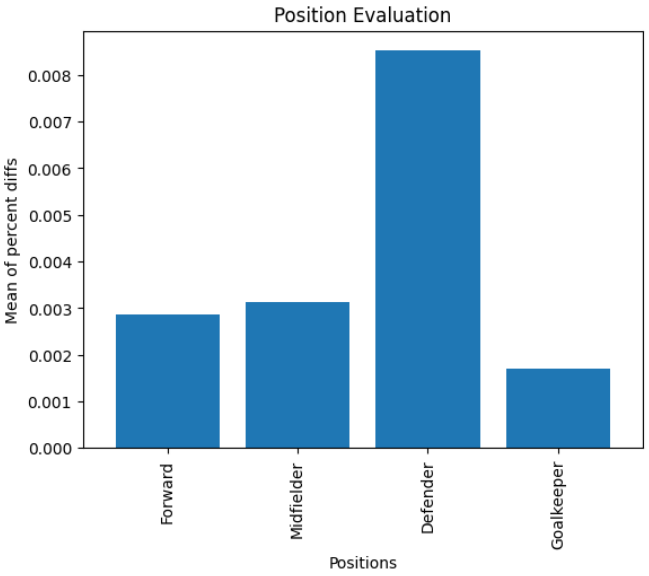
(d) Vyhodnocení shlukování pro útočníky

Obrázek 6.2: Vyhodnocení shlukování pro jednotlivé hráčské pozice

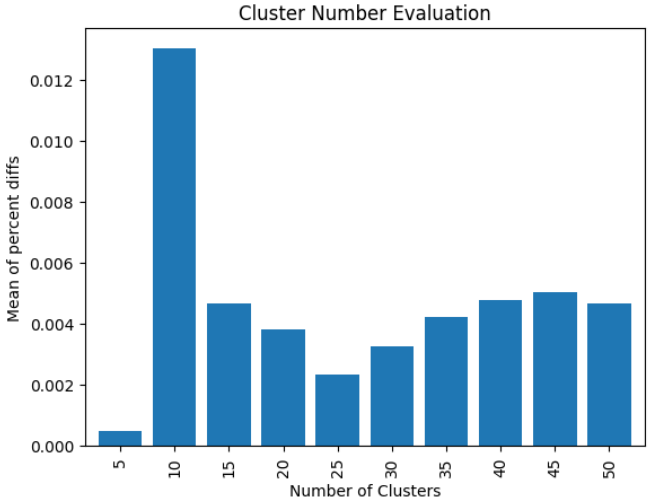
sadě nejvíc. Nutno říci, že mají nejlepší průměr více jak dvakrát větší než záložníci, kteří jsou druzí. Je to způsobeno tím, že skutečně malé procento běhů pro obránce je v mínusu (viz. graf 6.2b).



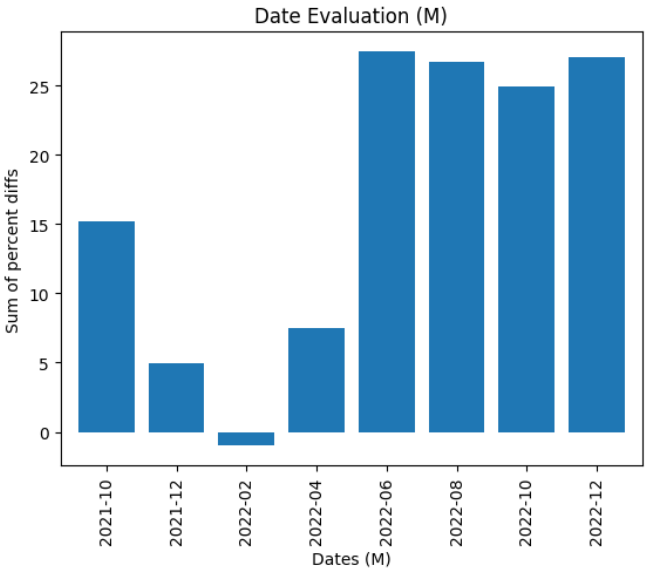
(a) Vyhodnocení skupin příznaků



(b) Vyhodnocení pozic



(c) Vyhodnocení k pro kmeans



(d) Vyhodnocení pro datumy

Obrázek 6.3: Vyhodnocení shlukování pro jednotlivé hráčské pozice

Závěr

Výsledkem práce je softwarové dílo, které zastřešuje všechny kroky strojového učení popsané v této práci, a experimentální činnost s využitím zmíněného programu. Jsme schopni, za předpokladu existence všech užitých datových zdrojů, stahovat nejaktuálnější data a zpracovávat je do tabulek. Následně tyto tabulky využíváme k vytvoření datových sad, které pomocí různých metod filtrujeme, trénujeme modely a výsledky vyhodnocujeme. Součástí práce je i vizualizace, vyhodnocení výsledků a jejich analýza. Práce poskytuje slibný základ pro další rozšíření ať už ve fázi získávání dat – obohacení dat o jiné kontexty jako třeba motivovanost týmu, psychické rozpoložení, počasí; vytváření příznaků – v této oblasti vidíme největší potenciál ke kreativitě z toho důvodu, že se stávajícími příznaky už sledované chybové metriky jen stěží klesaly; filtrace příznaků – přidání dalších metod jako například iterativní selekce; jiné modely a hyperparametry – zde je potenciál zejména v navrhnutí složitějšího MLP.

Pro oba definované problémy jsme ukázali vhodnost užití strojového učení, v případě regrese je to poražení průměrné predikce, v případě cen hráčů pak fakt, že při použití navrhnutého shlukovacího algoritmu průměrně karty narostly na ceně relativně vůči svému shluku.

Seznam použité literatury

Bangdiwala, M., Choudhari, R., Hegde, A. a Salunke, A. (2022). Using ml models to predict points in fantasy premier league. doi: 10.1109/ASIANCON55314.2022.9909447.

Bunker, R. a Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, **73**, 1285–1322. doi: 10.1613/jair.1.13509.

Hastie, T., Tibshirani, R. a Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition. URL `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`.

Horvat, T. a Job, J. (2020). The use of machine learning in sport outcome prediction: A review. **1**(2), 10–20.

Pappalardo, L. (2019). Soccer analytics: how data science is changing the "beautiful game".

Rodrigues, F. a Ângelo Pinto (2022). Prediction of football match results with machine learning. *Procedia Computer Science*, **204**, 463–470. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2022.08.057. URL `https://www.sciencedirect.com/science/article/pii/S1877050922007955`. International Conference on Industry Sciences and Computer Science Innovation.

Seznam obrázků

Seznam tabulek

Seznam použitých zkratek

- ETH = Ether, kryptoměna, token na platformě Ethereum, měna Sorare `https://ethereum.org/en/`
- NFT = Non-fungible-token, `https://en.wikipedia.org/wiki/Non-fungible_token`
- L<N>_<stat> = průměr za posledních N zápasu ve statistice stat (například L40_goals).
- L40 = pokud používáme pouze L40, znamená to průměr skóre. Tedy L40 je ekvivalentní s L40_totalScore.
- MSE = mean square error
- RMSE = root mean square error

A. Přílohy

A.1 Elektronická příloha

Byl přiložen vlastní kód, napsaný pro potřeby práce a grafy, které nebyly v práci přímo využity.

A.2 Prostor hyperparametrů pro každý model

1. **RidgeCV**
    - Žádné parametry k optimalizaci.
2. **MLPRegressor**
    - Hidden layer sizes: (100,), (100, 50), (100, 50, 10), (20, 20, 20, 20)
    - Activation: relu, tanh
    - Solver: adam, sgd
    - Alpha: 0.0001 to 0.01 (log scale)
    - Learning rate: constant, adaptive
3. **GradientBoostingRegressor**
    - Learning rate: 0.01 to 0.5 (log scale)
    - Number of estimators: 100 to 500
    - Max depth: 3 to 5
    - Min samples split: 2 to 10
    - Min samples leaf: 1 to 4
4. **SVR**
    - C: 1 to 10
    - Kernel: poly, rbf, sigmoid
5. **XGBRegressor**
    - Booster: gbtree, gblinear, dart
    - Eta: 0.01 to 0.5
    - Gamma: 0 to 5
    - Max depth: 3 to 10
6. **RandomForestRegressor**
    - Number of estimators: 100 to 1000
    - Max depth: None, 5, 20
    - Max features: sqrt, log2

- Min samples split: 2 to 10
- Min samples leaf: 2 to 4
- Bootstrap: True, False

A.3 Příznaky po ruční selekci

Tabulka A.1: Příznaky po ruční selekci pro brankáře

LS_L15_goalsConceded_home_mean
Průměrný počet inkasovaných gólů doma za posledních 15 zápasů
LS_L15_goalsConceded_away_mean
Průměrný počet inkasovaných gólů venku za posledních 15 zápasů
u_enemies_xGBuildup
Jak nebezpečné jsou akce, které soupeři vytváří, ale nekončí gólem
enemy_LS_L15_goalsScored_away_mean
Průměrný počet vstřelených gólů soupeřem venku za posledních 15 zápasů
gs_enemy_mean_pen_area_entries
Kolikrát se průměrně soupeř dostane do pokutového území
gs_enemy_mean_totalScore
Celkové průměrné skóre soupeře
enemy_LS_L15_avgWinOdds_all_mean
Průměrný kurz na výhru soupeře za posledních 15 zápasů
is_home
Zda je zápas doma nebo venku (1 pro doma, 0 pro venku)
enemy_LS_L15_goalsScored_home_mean
Průměrný počet vstřelených gólů soupeřem doma za posledních 15 zápasů
u_enemies_xGChain
Jak nebezpečné jsou gólové akce, které soupeři vytváří

Tabulka A.2: Příznaky po ruční selekci pro obránce

gs_player_mean_poss_won
Průměrný počet získaných míčů hráčem
LS_L15_goalsConceded_away_mean
u_enemies_xGBuildup
enemy_LS_L15_maxWinOdds_away_mean
enemy_LS_L40_goalsScored_home_mean
enemy_LS_L15_maxWinOdds_home_mean
gs_player_mean_mins_played
f_enemies_overall
Průměrné celkové ohodnocení ve FIFA pro soupeře
enemy_LS_L15_avgWinOdds_all_mean
is_home
gs_player_mean_interception_won
Průměrný počet přerušených přihrávek hráčem
L40
LS_L15_avgWinOdds_all_mean
u_teammates_xGChain
LS_L15_goalsConceded_home_mean
gs_player_std_allAroundScore
Standardní odchylka skóre pro hráče
gs_player_std_duel_won
Standardní odchylka počtu vyhraných soubojů hráčem
enemy_LS_L40_goalsScored_away_mean
u_teammates_xGBuildup
LS_L40_goalsConceded_home_mean
gs_ally_mean_successful_final_third_passes
Průměrný počet úspěšných přihrávek ve třetí části hřiště od spoluhráčů
enemy_LS_L15_goalsScored_away_mean
LS_L40_goalsConceded_away_mean
enemy_LS_L15_goalsScored_home_mean
u_enemies_xGChain
Jak nebezpečné jsou gólové akce, které soupeři vytváří

Tabulka A.3: Příznaky po ruční selekci pro útočníky

u_player_key_passes
f_teammates_movement_reactions
Průměrná reakce na pohyb míče ve FIFA pro spoluhráče
gs_player_mean_mins_played
LS_L40_shotsOnTarget_home_mean
u_teammates_xGChain
Jak nebezpečné jsou gólové akce, které spoluhráči vytváří
LS_L40_avgOver2.5_away_mean
enemy_LS_L15_goalsConceded_away_mean
gs_player_std_allAroundScore
LS_L40_goalsScored_home_mean
LS_L40_avgOver2.5_home_mean
Jaký je průměrný kurz na více než 2,5 gólu v zápase (L40)
f_teammates_overall
Průměrné celkové ohodnocení ve FIFA pro spoluhráče
LS_L40_goalsScored_away_mean
LS_L40_shotsOnTarget_away_mean
LS_L15_goalsScored_away_mean
L40
enemy_LS_L15_maxWinOdds_away_mean
Průměr z nejvyšších kurzů na výhru soupeře venku za posledních 15 zápasů
enemy_LS_L15_maxWinOdds_home_mean
Průměr z nejvyšších kurzů na výhru soupeře doma za posledních 15 zápasů
is_home
gs_player_mean_pen_area_entries
Kolikrát se průměrně hráč dostane do soupeřova pokutového území
u_player_xGBuildup
LS_L15_goalsScored_home_mean
enemy_LS_L15_goalsConceded_home_mean
Kolik gólů dostává soupeř doma průměrně (z posledních 15 zápasů)
gs_player_mean_successful_final_third_passes
u_teammates_xGBuildup
Jak nebezpečné jsou akce, které spoluhráči vytváří, ale nekončí gólem
u_player_xGChain
Jak nebezpečné jsou gólové akce, které hráč vytváří