FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Název:	Predikce výsledků zápasů v NHL
Student:	Filip Kojan
Vedoucí:	Ing. Karel Klouda, Ph.D.
Studijní program:	Informatika
Studijní obor:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	Do konce letního semestru 2019/20

Pokyny pro vypracování

1) Proveďte rešerši zdrojů dat o zápasech a hráčích NHL. Zaměřte se na dostupnost tzv. moderních statistik (CORSI apod.).
2) Proveďte rešerši známých metod používaných pro predikce výsledků zápasů kolektivních sportů.
3) Ze získaných dat vytvořte vhodné příznaky a na nich experimentálně porovnejte vybrané metody predikce výsledků zápasů NHL.
4) Výsledky porovnejte také s predikcemi sázkových kanceláří.

Seznam odborné literatury

Dodá vedoucí práce.

Ing. Karel Klouda, Ph.D.
vedoucí katedry

doc. RNDr. Ing. Marcel Jiřina, Ph.D.
děkan

V Praze dne 23. ledna 2019

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Bakalářská práce

Predikce výsledků zápasů v NHL

Filip Kojan

Katedra aplikované matematiky
Vedoucí práce: Ing. Karel Klouda, Ph.D.

12. května 2019

Poděkování

Rád bych poděkoval vedoucímu mé práce Ing. Karlu Kloudovi, Ph.D. za ochotu, trpělivost, věnovaný čas a cenné rady, bez kterých by tato práce nemohla vzniknout. Dále bych rád poděkoval rodině, která mi poskytla podporu jak při tvorbě bakalářské práce, tak při studiu na ČVUT.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů. V souladu s ust. § 46 odst. 6 tohoto zákona tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla, avšak pouze k nevýdělečným účelům. Toto oprávnění je časově, teritoriálně i množstevně neomezené.

V Praze dne 12. května 2019 ……

Odkaz na tuto práci

Kojan, Filip. *Predikce výsledků zápasů v NHL*. Bakalářská práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2019.

Abstrakt

Předmětem této práce je prozkoumání zdrojů dat o hráčích a zápasech v hokejové NHL, moderních statistických metod používaných k vyhodnocení kvality týmů a hráčů a využití těchto informací k predikci výsledků zápasů v NHL. Použity jsou různé klasifikační modely a je porovnána jejich přesnost. Dále jsou výsledky predikcí porovnány s predikcemi sázkových kanceláří.

Klíčová slova Predikce, NHL, Lední hokej, Moderní statistiky, Corsi, Fenwick, PDO, Klasifikace, Python, Strojové učení

Abstract

The goal of this thesis is to explore data sources about players and matches in NHL and about modern statistic methods, which are used for evaluating quality of teams and players and possibilities of using these informations for predicting results of NHL matches. Various classification models of machine learning are used and their predictive ability is compared. The results of predictions are compared to bookmaker predictions.

Keywords Prediction, NHL, Ice hockey, Modern statistics, Corsi, Fenwick, PDO, Classification, Python, Machine learning

Obsah

Seznam obrázků

Seznam tabulek

Seznam výpisů

Úvod

Americká NHL (National Hockey League, česky Národní hokejová liga) je bezpochyby jednou z nejprestižnějších hokejových soutěží na světě. Přesto se jí nedostává tolik pozornosti v oblasti predikcí výsledků, jako ostatním, hlavním, americkým, národním ligám, jako jsou NFL (National Football League, česky Národní fotbalová liga) a NBA (National Basketball Association, česky Národní basketbalová asociace). To může být způsobeno i tím, že je náročné analyzovat hokejovou hru jako takovou, a také nižším počtem příležitostí ke skórování (oproti zmiňovaným NFL a NBA, kde je závěrečné skóre zápasu často o řád vyšší než právě v NHL). V posledních letech se rozšiřuje množství dat zaznamenávaných o jednotlivých akcích na ledě. S těmito rozšířeními rostou i možnosti analýzy hokejové hry a tím pádem se zvětšuje prostor pro predikce (nejenom) výsledků zápasů.

Tématem této práce jsou predikce výsledků zápasů v NHL ve smyslu výhry domácího týmu, nebo výhry hostů. Tyto predikce mohou pomoci sázkařům, kteří mohou tyto predikce využít ke zvýšení zisku při sázení, samotným sázkovým kancelářím, které mohou na základě predikcí upravovat své sázkové kurzy, a tím maximalizovat zisky.

Motivací k výběru tématu bylo převážně to, že existuje relativně malé množství prací, které se zabývají predikcí výsledků hokejových zápasů, jak bylo zmíněno výše v úvodu. Dále i poměrně malá přesnost predikcí u sázkových kanceláří a tudíž velký potenciál zisku v případě úspěšných predikcí.

Práce je zaměřena na získání dostupných dat k jednotlivým zápasům NHL, vybrání vhodných příznaků, jejich transformací do formátu vhodného k použití v klasifikačních modelech a následné vyhodnocení přesnosti predikce.

V první části bude čtenář seznámen se základními informacemi o průběhu sezóny v NHL a s tzv. moderními statistikami používanými v ledním hokeji. V další části se práce zabývá dostupností zmíněných dat a jejich popisem. Ve třetí části se práce zabývá výběrem vhodných příznaků, které mají největší vliv na výsledek zápasu. V následující části se práce zaobírá výběrem vhodných

klasifikačních modelů a v závěrečné části se práce věnuje porovnání výsledků predikcí s predikcemi výsledků sázkových kanceláří.

Jak bylo zmíněno výše v úvodu, přesnost sázkových kanceláří není úplně vysoká, přesněji se jedná o 60,76 % (pokud bereme v potaz pouze predikci výhra/prohra, tedy stejný typ predikce, které se věnuje tato práce). Tato úspěšnost byla vypočítána ze zápasů od sezóny 2011/2012 (včetně), až po 11. 2. 2019 a byly započteny všechny proběhlé zápasy tzv. základní části. Celkově se jedná o 9000 utkání. Této úspěšnosti se práce pokouší vyrovnat, neboť sázkové kanceláře by měly mít nejpřesnější predikce, už jen z jejich vlastní podstaty.

Cíl práce

Cílem teoretické části práce je prozkoumat zdroje dat o zápasech NHL se zaměřením na dostupnost moderních, hokejových statistik, jako jsou Corsi nebo Fenwick. Dalším cílem teoretické části je rešerše existujících metod pro predikci výsledků kolektivních sportů, a rešerše zdrojů a dat dostupných k proběhlým zápasům NHL.

Cílem praktické části práce je získat data ze zdrojů prozkoumaných v teoretické části a ze získaných dat vytvořit příznaky vhodné pro použití v klasifikačních modelech. Nad těmito příznaky experimentovat a tím získat co nejlepší klasifikační přesnost vybraných klasifikačních modelů a na závěr porovnat výsledky predikcí těchto modelů s predikcemi sázkových kanceláří.

Kapitola 1

Lední hokej

1.1 NHL

NHL je profesionální kanadsko-americká hokejová soutěž, která je považována za jednu z nejprestižnějších hokejových lig světa. Historie této ligy sahá až do roku 1917, kdy vznikla z kanadské NHA(National Hockey Association, česky Národní hokejová asociace). [1]

V současné době hraje NHL 31 týmů, přičemž nejmladší tým je Vegas Golden Knights, který se soutěže účastnil poprvé v sezóně 2017/2018. [2]

Týmy v soutěži jsou rozděleny do dvou konferencí, Východní a Západní, a v každé konferenci dále ještě do dvou divizí. Divize ve Východní konferenci jsou divize Atlantická, obsahující 8 týmů a Metropolitní, také s 8 týmy. V Západní konferenci jsou divize Pacifická, obsahující 8 týmů a Centrální, ve které hraje týmů 7.

Sezóna v NHL se dělí na dvě hlavní části. První částí je tzv. základní část a druhou částí je playoff. Základní část začíná zpravidla začátkem října a končí začátkem dubna. Na základní část navazuje část vyřazovací, tzv. playoff, kam postupuje 16 nejlepších týmů ze základní části. Soutěž zpravidla končí začátkem června.

V základní části hraje každý z týmů 82 zápasů, z nichž je 41 na domácím ledě a 41 na ledě soupeře. Týmy spolu hrají napříč celou soutěží, tedy nejenom v rámci divize nebo konference. Výjimkou byla sezóna 2012/2013, kdy proběhl tzv. NHL lockout (výluka). Zde se začátek sezóny posunul až na 19. února. Důvodem byl konec kolektivní smlouvy mezi NHLPA (National Hockey League Player Association, česky Hráčská asociace NHL) a samotnou NHL. Problém byla neschopnost domluvit nové podmínky této smlouvy. [3] Každý z týmů odehrál v této sezóně pouze 48 zápasů místo obvyklých 82.

Do playoff postupují tři nejlepší týmy (týmy s nejvíce body) z každé divize. Těchto 12 týmů je doplněno dalšími čtyřmi týmy, dvěma z každé konference bez ohledu na divizi. Je tedy možné, aby v playoff skončilo 5 týmů z jedné

divize a pouze 3 týmy z druhé v každé konferenci. Tyto čtyři týmy jsou vybrány na základě nejvyššího počtu bodů a budou držiteli tzv. divokých karet.

Playoff se hraje jako vyřazovací turnaj, kde týmy hrají na čtyři vítězná utkání. Obě dvě konference mají nezávislá playoff, tudíž v playoff není možné, aby spolu hrály týmy z opačných konferencí. Toto pravidlo platí až do finále, kdy naopak proti sobě vždy hrají týmy z opačných konferencí. Struktura playoff je zobrazena na obrázku 1.1.

Obrázek 1.1: Struktura playoff [4]

1.2 Moderní hokejové statistiky

Tato práce se bude věnovat třem (pravděpodobně) nejznámějším moderním statistikám a těmi jsou Corsi, Fenwick a PDO. Corsi a Fenwick jsou statistiky zaměřující se na různé druhy střel a jejich počty. Obě tyto statistiky jsou aplikovatelné jak na jednotlivé hráče, tak i na celé týmy. Ani jedna ze zmíněných statistik nebere žádným způsobem v potaz góly, řeší pouze střely a nezáleží na tom, jestli z dané střely padla či nepadla branka. Tím se liší od třetí statistiky PDO. Tato statistika je součtem úspěšnosti střelby daného týmu a úspěšnosti brankáře týmu.

Výhodou těchto statistik je také to, že jsou relativně jednoduše získatelné z dat, která jsou k danému zápasu k dispozici. To je rozdíl od další z moderních statistik, kterou jsou tzv. *zone starts*. Jedná se o statistiku, která analyzuje, kdy je daný hráč nasazován na led, jestli spíše při ofenzivní hře, tedy když jeho vlastní tým útočí, nebo při defenzivní hře, kdy je vlastní tým v obranné

třetině. Tato data je poněkud složitější získat (nikoliv však nemožné), proto se práce této statistice nevěnuje.

1.2.1 Corsi

První statistikou, která je v práci využita, je statistika Corsi. Statistika je pojmenována po trenérovi brankářů z týmu Buffalo Sabres, Jimovi Corsim [5]. Jedná se o statistiku, která je velmi podobná +/- bodům. Corsi se dá použít jak pro celé týmy, tak i pro jednotlivé hráče. Tato statistika se snaží vyhnout problému +/- statistiky, která počítá rozdíl mezi vstřelenými a inkasovanými góly, zatímco byl konkrétní hráč na ledě (při hře 5 na 5, tedy nebere v potaz góly v přesilových hrách a v oslabení). Problém +/- statistiky je, že gólů padá relativně málo, tudíž může být ovlivněna smůlou či štěstím daného hráče. Tomuto se snaží Corsi předejít tím, že nebere v potaz pouze góly, ale všechny střely. Ve chvíli, kdy jeden z týmů vystřelí, ať už na soupeřovu branku, mimo ní, nebo je střela zablokována bránicím hráčem, je hráčům tohoto týmu, kteří jsou zrovna na ledě, přičten jeden bod do statistiky *Corsi for*, dále jen *CF*, a hráčům druhého týmu, je přičten bod do statistiky *Corsi against*, dále jen *CA*. Celkové Corsi, dále jen *C*, daného hráče je poté počítáno jako rozdíl *CF* a *CA*. Z těchto statistik lze také spočítat tzv. *Corsi for percentage*, dále jen *C%* a to jako $C\% = CF/(CF + CA)$.

1.2.2 Fenwick

Fenwick je velmi podobná statistika jako *Corsi*. Jediným rozdílem je, že v této statistice se neberou v potaz střely zblokované bránicím hráčem. Statistika byla pojmenována podle blogera z webu Battle of Alberta a fanouška týmu Calgary Flames, Matta Fenwicka. [6] Proč využívat Fenwick statistiku namísto (nebo zároveň s) Corsi shrnul právě Matt Fenwick.

„My argument is basically:
1. The whole (or perhaps best) use of Corsi is to have objective figures that can be used as a proxy for scoring chances (what else are you using it for?).
2. A shot that is blocked is either a) not a scoring chance at all, or b) on average from a worse scoring area than shots/posts/missed shots.
Yes it affects the „sample size“ but that only means anything if what you are sampling is relevant to what you are trying to represent. You could include Penalties Drawn (or hell, faceoff wins/losses!), but I'm not sure that the connection to a scoring chance is as obvious as for shots/posts etc. That's my opinion anyway, take it or leave it.“

- Matt Fenwick, Nov. 22nd, 2007. [6]

1.2.3 PDO

PDO je často nazývaná jako *„statistika štěstí“* nebo *„index štěstí“*. [7] Jedná se o statistiku, kterou NHL pojmenovává jako *SPSV%*. [8] *PDO* je statistika,

která je jednoduchým součtem střelecké úspěšnosti a úspěšnosti brankářských zákroků daného týmu. Vzhledem k tomu, že se jedná o inverzní údaje (úspěšnost střelby jednoho týmu je inverzní k úspěšnosti brankáře druhého týmu) je v každém zápase průměr *PDO* hrajících týmů 100. Stejný průměr těchto hodnot je i v celé soutěži.

Tato statistika vychází z předpokladu, že týmy hrající stejnou soutěž (v našem případě NHL), jsou vyrovnané. Tudíž tým, který má *PDO* výrazně přes 100, se považuje za tým s větším štěstím a tým, který má *PDO* výrazně pod 100, se považuje za tým, který doprovází spíše smůla. Tím pádem, se zvyšujícím se počtem odehraných zápasů, by toto číslo mělo konvergovat k průměru soutěže, tedy 100.

Kapitola 2

Analýza zdrojů

Tato kapitola se zabývá zdroji, které poskytují požadované informace. Jedná se o záznamy o zápasech a sázkové kurzy použitelné pro porovnání výsledků.

2.1 Webové stránky soutěže NHL

Web NHL je velmi rozsáhlým zdrojem informací o celé soutěži. Jsou zde záznamy o zápasech od sezóny 1917/1918 až po nejaktuálnější zápasy. Vzhledem k vývoji hokejové hry budou v práci využity pouze záznamy ze sezón 2011/2012 a novějších. Dalším důvodem pro použití pouze těchto „novějších" dat je to, že podrobné statistiky o jednotlivých střelách byly zaznamenávány až od roku 2010. Z dřívějších záznamů by tedy nebylo možné počítat statistiky Corsi a Fenwick. Další nespornou výhodou tohoto oficiálního zdroje je to, že poskytuje REST API (Application Programming Interface). Bohužel toto API není oficiálně zdokumentované, ale existuje dokumentace vytvořená jedním z fanoušků na webowé adrese `https://gitlab.com/dword4/nhlapi`. Výchozí adresa pro REST API je `https://statsapi.web.nhl.com/api/v1`. Toto API poskytuje veškerá data ve formátu JSON, tudíž jsou tato data jednoduše strojově čitelná a zpracovatelná.

Z tohoto zdroje lze získat nepřeberné množství různých statistik a informací. Pro potřeby práce jsou nejdůležitější detailní informace o průběhu každého zápasu na adrese `/game/{ID}/feed/live`, kde `{ID}` je unikátní identifikátor zápasu. Na této adrese najdeme, pro každý zápas, velmi detailní popis jeho průběhu. Detailnímu popisu těchto dat se věnuje kapitola 3.

Velmi důležitou součástí vstupních dat jsou i data o aktuální pozici v tabulce. Tato data jsou taktéž poskytována pomocí REST API a to na adrese `/standings?date=2018-12-19` (v tomto případě k datu 19. 12. 2018). Záznamy o pozici v tabulce zde nalezneme k jakémukoliv datu. Najdeme zde informace o počtu bodů, počtu vstřelených a inkasovaných branek, pozici v tabulce v rámci divize, konference i celé soutěže, počtu odehraných zápasů

a mnoho dalšího. Detailní popis těchto dat čtenář opět nalezne v kapitole 3.

Na webu `nhl.com` také nalezneme záznamy o střídání v jednotlivých zápasech. Jsou ovšem poskytovány na jiné adrese, než předchozí informace, ačkoliv se stále jedná o stejné webové stránky. Záznamy o střídání lze dohledat na adrese `http://www.nhl.com/stats/rest/shiftcharts?cayenneExp=gameId=`{ID}, kde {ID} je opět unikátní identifikátor zápasu. Tato data, spolu s informacemi o průběhu zápasu, jsou klíčová pro výpočet moderních statistik jako jsou Corsi a Fenwick.

S odkazem na *Terms of Service, bod č. 7*, na webových stránkách NHL, na adrese `https://www.nhl.com/info/terms-of-service`, lze říci, že veškeré informace dostupné na tomto webu, jsou majetkem NHL nebo třetích stran a je možné je využít pouze pro nekomerční účely, což je v souladu s využitím těchto dat v této práci.

2.2 Web livesport.cz

Dalším typem dat, která budou pro potřeby práce nutná, jsou historické sázkové kurzy. Tyto kurzy bylo relativně náročné na internetu dohledat v úplné podobě. Nakonec se jako nejjednodušší varianta ukázalo použití kurzů z webové stránky `livesport.cz`. Zde se nachází archiv zápasů NHL a ke každému zápasu jsou zde k dispozici i sázkové kurzy. Tyto kurzy byly z webu staženy napodobením GET požadavku, který posílá zmíněná webová stránka do svých vlastních zdrojů.

Výňatek z podmínek užití na webu `livesport.cz`: *Návštěvníci nejsou bez předchozího písemného povolení ze strany poskytovatele oprávněni kopírovat, upravovat, distribuovat, přenášet, zobrazovat, reprodukovat, nahrávat, stahovat ani jinak používat obsah aplikace ani do něj nijak zasahovat nebo měnit cokoli z jejího obsahu.* [9] Toto povolení bylo ze strany `livesport.cz` uděleno a je dostupné na přiloženém CD.

Kapitola 3

Popis dat

Tato kapitola se zabývá popisem dat stažených pro potřeby práce. Veškerá data byla stažena do lokálního počítače, aby byla dostupná neustále, bez závislosti na dostupnosti originálního zdroje. Celkem se jednalo o přibližně 6 GB dat. Vzhledem k této, relativně malé, velikosti, bylo možné pro uložení využít pevný disk v počítači. Všechna data byla stažena pomocí skriptů v programovacím jazyce Python.

3.1 Záznam o zápase

Data ze záznamů o zápasech byla stažena z webové adresy *https://statsapi.web.nhl.com/api/v1/game/{ID}/feed/live*, kde {ID} je unikátní identifikátor daného zápasu. Tyto záznamy jsou ve formátu JSON, tudíž jsou jednoduše počítačově zpracovatelné. Záznam o zápase poskytuje velké množství informací. Záznam o každém zápase má přibližnou velikost mezi 350kB a 500kB a i to může napovědět o množství dat v něm obsažených. Záznam je rozdělen do několika úrovní/vrstev. Na každé úrovni obsahuje záznam několik klíčů. Jednotlivé klíče jsou popsány v následujícím seznamu.

- *gamePk* - Jedná se o unikátní identifikátor zápasu, který se skládá z 10 číslic. První 4 číslice odkazují na sezónu, ve které daný zápas proběhl, např. 2011 odkazuje na sezónu 2011/2012. Další dvě číslice jsou buď 01, 02, 03, nebo 04. Ty poskytují informaci o typu zápasu. 01 pro preseason (příprava), 02 pro regular season (základní část), 03 pro playoff (vyřazovací část) a 04 pro all-star hry (zápasy hvězd). A poslední čtveřice číslic specifikuje číslo hry (od 0001 do 1271 pro sezóny s 31 týmy a od 0001 do 1230 pro sezóny s 30 týmy). [10] Výjimkou je sezóna 2012/2013, kdy bylo odehráno pouze 720 zápasu, což bylo způsobeno výlukou v NHL.

- *gameData* - Zde jsou (mimo jiné) uloženy záznamy o jednotlivých týmech, hráčích, stadionu, kde se zápas hraje nebo např. o datu zápasu.

Součástí této části dat je:

  - *teams* - Jedná se o záznamy o hrajících týmech, který z týmů je domácí a který hostující, unikátní identifikátory týmů, názvy, zkratky, města původu, domácí stadiony a další.
  - *players* - Záznam o každém hráči, který byl pro daný zápas zaregistrovaný. Záznam obsahuje unikátní identifikátor hráče, jeho jméno, národnost, datum narození, číslo dresu, váhu, výšku, tým za který momentálně hraje a pozici, jakou v týmu zastává.
  - *venue* - Informace o stadionu, na kterém zápas probíhá, obsahuje pouze název stadionu, jeho identifikátor a link odkazující na REST API na webu `nhl.com`.
  - *datetime* - Záznam obsahující pouze dvě informace. Datum a čas začátku a konce zápasu.

- *liveData* - Zde najdeme již informace o samotném průběhu daného zápasu. Jsou zde informace o vyhlášených nejlepších hráčích, o jednotlivých akcích na ledě nebo tzv. *boxscore*.
  - *decisions* - Záznam obsahující informace o vyhlášených hvězdách zápasu.
  - *boxscore* - Zde jsou uloženy statistiky, týkající se daného zápasu, pro každého hráče. Tyto statistiky obsahují informace např. o tom, kolik každý hráč v zápase vstřelil branek, zablokoval střel, kolikrát vystřelil, kolik času na ledě strávil, čas na ledě při přesilovce, počet trestných minut, kolik vyhrál a prohrál vhazování. Bohužel zde nejsou moderní statistiky jako je Corsi nebo Fenwick. Tyto statistiky budou dopočítány „ručně“ z dostupných dat.
  - *plays* - Záznamy o akcích na ledě. Pro účely práce se jedná o jedny z nejdůležitějších informací, neboť zde najdeme záznamy o každé střele, ať už na branku, mimo branku nebo zblokované, trestech (potažmo přesilových hrách), vhazování nebo tzv. hitech. Jedná se o velmi podrobné záznamy, např. u střel je zde zaznamenán i typ střely, jako je *wrist, slap, snap, backhand, tip-in*. Každý záznam o akci je doplněn, časem a třetinou ve které nastala, pozicí na ledě zadanou pomocí *X, Y* souřadnic, identifikátory hráčů, kteří se dané akce účastnili, jednoduchým popisem dané situace a aktuálním stavem skóre zápasu. Jak záznam o akci na ledě vypadá, lze vidět ve výpisu 3.1.

```
{ "players":[ {
      "player":{
        "id":8467329,
        "fullName":"Vincent Lecavalier",
        "link":"/api/v1/people/8467329" },
      "playerType":"Blocker" },{
      "player":{
        "id":8470187,
        "fullName":"Johnny Boychuk",
        "link":"/api/v1/people/8470187" },
      "playerType":"Shooter" } ],
  "result":{
    "event":"Blocked Shot",
    "eventCode":"BOS6",
    "eventTypeId":"BLOCKED_SHOT",
    "description":"Vincent Lecavalier blocked
    shot from Johnny Boychuk" },
  "about":{
    "eventIdx":6,
    "eventId":6,
    "period":1,
    "periodType":"REGULAR",
    "ordinalNum":"1st",
    "periodTime":"00:44",
    "periodTimeRemaining":"19:16",
    "dateTime":"2011-10-08T23:17:27Z",
    "goals":{
      "away":0,
      "home":0 } },
  "coordinates":{
    "x":65.0,
    "y":-6.0 },
  "team":{
    "id":14,
    "name":"Tampa Bay Lightning",
    "link":"/api/v1/teams/14",
    "triCode":"TBL" }}
```


Výpis 3.1: Ukázka záznamu o jedné akci

3.2 Záznamy o střídání

Jak bylo zmíněno výše, v kapitole 2, data se záznamy o střídání pochází také z webu `nhl.com`, ale byla stažena z jiné adresy, než data o jednotlivých zápasech. Následuje ukázka 3.2 výpisu jednoho střídání z utkání s ID *2011020001*. Konkrétní webová adresa pro tento zápas je: `http://www.nhl.com/stats/rest/shiftcharts?cayenneExp=gameId=2011020001`

```
{'data': [...,{'detailCode': 0,
               'duration': '00:34',
               'endTime': '13:11',
               'eventDescription': None,
               'eventDetails': None,
               'eventNumber': None,
               'firstName': 'Brad',
               'gameId': 2011020001,
               'hexValue': '#111111',
               'lastName': 'Marchand',
               'period': 2,
               'playerId': 8473419,
               'shiftNumber': 13,
               'startTime': '12:37',
               'teamAbbrev': 'BOS',
               'teamId': 6,
               'teamName': 'Boston Bruins',
               'typeCode': 517},...],
 'total': 801}
```

Výpis 3.2: Ukázka záznamu o střídání

V záznamu vidíme seznam *data*, který obsahuje záznamy o jednotlivých střídání. Pro nás podstatné informace jsou:

- *gameID* - Unikátní identifikátor zápasu, pomocí kterého můžeme přiřadit daný záznam o střídání ke konkrétnímu zápasu.
- *playerID* - Unikátní identifikátor hráče, který nám určí, kterého hráče se dané střídání týká.
- *startTime, endTime, duration* - Záznamy, kdy daný hráč nastoupil na led, kdy byl opět vystřídán a rozdíl těchto dvou časů.
- *period* - Číslo třetiny, abychom správně určili čas střídání, neboť časy jsou uváděny v rámci třetiny (tedy 00:00-20:00) a ne v rámci celého utkání.

Dále zde nalezneme záznam *total* na stejné úrovni, jako seznam *data*. Jedná se o celkový počet všech střídání v daném zápase. Ostatní záznamy pro nás nejsou podstatné, neboť v práci nebudou využity.

3.3 Umístění v tabulce

Další, neméně podstatnou součástí dat, je umístění týmů v tabulce. Na webu NHL jsou tyto záznamy dostupné ke každému dni. Opět je zde využitá standardní webová adresa pro REST API `https://statsapi.web.nhl.com/api/v1/` s tím, že za lomítkem adresa pokračuje `standings?date=2019-02-09`. Tato adresa vrátí pozici všech týmů i s dalšími detaily ke dni 9. 2. 2019. Datum může být nahrazeno libovolným jiným validním datem. V následující ukázce 3.3 můžeme vidět výpis z tabulky týmu „New York Islanders“.


```
{ "team":{
    "id":2,
    "name":"New York Islanders",
    "link":"/api/v1/teams/2"
  },
  "leagueRecord":{
    "wins":17,
    "losses":12,
    "ot":4,
    "type":"league"
  },
  "goalsAgainst":93,
  "goalsScored":96,
  "points":38,
  "divisionRank":"3",
  "conferenceRank":"8",
  "leagueRank":"16",
  "wildCardRank":"0",
  "row":15,
  "gamesPlayed":33,
  "streak":{
    "streakType":"wins",
    "streakNumber":3,
    "streakCode":"W3"
  },
  "lastUpdated":"2019-03-21T01:13:25Z"}
```


Výpis 3.3: Záznam z tabulky

V záznamu nalezneme následující údaje.

- *team* - Identifikátor a celý název daného týmu.
- *leagueRecord* - Počet vyhraných, prohraných a remizovaných zápasů.
- *goalsAgainst* - Počet inkasovaných branek.
- *goalsScored* - Počet vstřelených branek.
- *points* - Počet bodů týmu.
- *divisionRank* - Pozice týmu v jeho divizi.
- *conferenceRank* - Pozice týmu v jeho konferenci.
- *leagueRank* - Pozice týmu v celé soutěži napříč divizemi i konferencemi.
- *wildCardRank* - Pozice v tabulce k získání divoké karty; 0 v případě, že se tým nachází na příčkách, ze kterých týmy postupují do playoff i bez divoké karty.
- *row* - Počet zápasů vyhraných ve standardní době nebo v prodloužení. Tedy celkový počet výher týmu bez výher na samostatné nájezdy.
- *gamesPlayed* - Počet odehraných zápasů v sezóně.
- *streak* - Počet zápasů v řadě, které daný tým vyhrál, případně prohrál.

Tyto informace jsou využity při tvorbě výsledného datasetu.

Kapitola 4

Práce s daty

Tato kapitola se zabývá stažením a zpracováním dat do formy vhodné k použití v klasifikačních modelech.

4.1 Použité nástroje

Všechna data použitá v této práci byla stažena i zpracovávána pomocí skriptů v programovacím jazyce Python, verze 3.6.7.

Pro práci s Python skripty byla využita open-source webová aplikace Jupyter notebook. Nejedná se o plnohodnotné IDE (Integrated Development Environment, česky Vývojové prostředí) jako je například PyCharm (vývojové prostředí pro Python od společnosti JetBrains). Jupyter například neprovádí kontrolu kódu v reálném čase, ani neumožňuje debuggovat kód. Důvodem, proč byl zvolen pro práci právě Jupyter notebook, je jeho jednoduchost a možnost spouštění kódu v jednotlivých blocích. Tato funkcionalita byla užitečná převážně při práci s klasifikačními modely, kdy nebylo nutné spouštět celý skript znovu (včetně učení modelu, časově nejnáročnější části), ale stačilo znovu spustit pouze jednotlivé bloky a například přidat nebo odebrat některé příznaky. Další výhodou Jupyter noteboooku je možnost zobrazovat výstupy kódu (např. tabulky, grafy atp.) přímo pod danými buňkami (bloky), což pomáhá přehlednosti kódu.

4.1.1 Dostupný hardware

Veškeré operace nad daty (stahování dat, tvorba datasetu, předpříprava dat, učení modelů...) byly prováděny na počítači (notebooku) s následujícími parametry:

```
Výrobce a typ notebooku: Lenovo Y50-70
Procesor: Intel Core i5 4210H
RAM: 8 GB
Pevný disk: Samsung SSD 840 Pro Series, 256 GB
Grafické karty: NVidia GeForce GTX 860M, 4 GB,
                Intel HD Graphics 4600
Operační systém: Windows 10 Education
```

4.1.2 Použité knihovny

Python je oblíbený i z důvodu velkého množství existujících knihoven. Nejdůležitější knihovny použité v této práci jsou popsány v následujícím seznamu.

- *scikit-learn* - Knihovna určená pro strojové učení. Obsahuje velké množství jak klasifikačních, tak i regresních modelů. Dále obsahuje metody pro předzpracování dat a metody pro výběr modelů a jejich parametrů.
- *pandas* - Open-source knihovna, která umožňuje jednoduchou a rychlou práci s datovými strukturami. V této práci byla tato knihovna použita hlavně pro práci s daty uleženými v tzv. *DataFrame*.
- *urllib* - Knihovna, která byla použita při stahování dat. Umožňuje posílat požadavky na různé *URL* (Uniform Resource Locator, česky jednotná adresa zdroje) a kontrolovat navrácená data.
- *json* - Knihovna pro práci s datovým formátem JSON.

4.2 Stahování dat

Stahování dat probíhalo ve dvou fázích. První fází bylo stažení dat z webu `nhl.com`. Vzhledem k REST API, které web poskytuje, bylo stažení relativně jednoduché. Ale vzhledem k počtu zápasů (9000 utkání) trvalo stažení dat několik minut. Data k jednotlivým zápasům a střídání byla ukládána do samostatných souborů (jeden zápas - jeden soubor). Data k postavení v tabulce byla stahována a ukládána do souborů ke každému dni (jedno datum - jeden soubor). Skripty pro stažení záznamů o zápase, střídání a postavení v tabulce jsou k dispozici na přiloženém CD pod názvy *games.ipynb*, *shifts.ipynb* a *standings.ipynb*.

Druhou fází bylo stažení dat z webu `livesport.cz`. Tento web neposkytuje žádné API, tudíž bylo nutné najít jiný způsob stažení dat. Jako nejefektivnější způsob se ukázalo „ruční“ stažení archivu jednotlivých sezón v NHL ve formátu HTML. Poté bylo nutné použít vlastní parser, který ze záznamů získal unikátní identifikátor každého zápasu, který web `livesport.cz` používá. Tento identifikátor byl následně použit, při napodobení GET požadavku, který `livesport.cz` používá pro dotazování do vlastních informačních zdrojů. Skript

pro parsování dat ze stažených HTML souborů, následnou tvorbu a odeslání GET požadavku je k dispozici na přiloženém CD pod názvem *kurzy web scraping.ipynb*

4.3 Tvorba statistik

Po stažení záznamů o jednotlivých zápasech bylo nutné vytvořit statistiky jednotlivých hráčů a týmů. Tomu se věnuje skript s názvem *stats calculating.ipynb*. Tento skript projde každý záznam o utkání a vytvoří soubor se statistikami jednotlivých hráčů a týmů. Tyto statistiky jsou vztaženy vždy k období od začátku sezóny až po datum, kdy se dané utkání odehrálo. Statistiky pro hráče a týmy byly vždy ukládány do totožného souboru. Statistiky brankářů nebyly ukládány mezi ostatní hráče, ale samostatně, neboť u brankářů se zaznamenávají diametrálně odlišné informace.

V tomto skriptu byly počítány následující statistiky.

- *pro celé týmy:*
  - *Střely na bránu*
  - *Zásahy brankářů*
  - *Inkasované branky*
  - *Obdržené branky*
  - *Corsi for*
  - *Corsi against*
  - *Corsi* - Rozdíl předchozích dvou údajů.
  - *Corsi percentage* - Popis v kapitole 1.2.1.
  - *USAT* - Počet nezblokovaných střel.
  - *Fenwick* - Popis v kapitole 1.2.2.
  - *PIM* - Obdržené trestné minuty.
  - *Úspěšnost střelby*
  - *Úspěšnost zásahů brankářů*
  - *PDO* - Popis v kapitole 1.2.3.
  - *Počet hodin od posledního zápasu*
- *pro jednotlivé hráče:*
  - *Corsi for*
  - *Corsi against*
  - *Corsi* - Rozdíl předchozích dvou údajů.

  - *Corsi percentage* - Popis v kapitole 1.2.1.

- *pro brankáře:*
  - *Inkasované branky*
  - *Úspěšné zásahy*
  - *Úspěšnost zásahů*

Nejnáročnější statistikou na výpočet, byla statistika *Corsi* pro jednotlivé hráče. Hlavním důvodem náročnosti výpočtu bylo, že pro každou střelu v zápase (které se do statistiky *Corsi* započítávají všechny) bylo nutné zjistit, kteří hráči byli v době střely na ledě a těmto hráčům přičíst bod do statistiky *Corsi for*, potažmo *Corsi against*. Tento údaj se v datech poskytnutých webem `nhl.com` přímo nevyskytuje a proto bylo nutné tuto informaci získat z dat o střídání. Zároveň bylo nutné identifikovat střely v přesilových hrách, neboť ty do statistiky nejsou započítávány. Výsledkem bylo, že tvorba těchto statistik trvala přibližně 14 minut (pro zmíněných 9000 zápasů a na zmíněném hardwaru).

4.4 Tvorba datasetu

V této části se práce zabývá tvorbou datasetu, který je využíván v klasifikačních modelech. Popis statistik dostupných pro tvorbu datasetu je k dispozici v kapitole 4.3. Finální dataset je tvořen 9000 řádky (zápasy) a 88 sloupci (příznaky). Každý řádek datasetu reprezentuje právě jedno utkání a jsou v něm statistiky týmů před proběhnutím daného zápasu. V následujícím seznamu jsou popsány jednotlivé sloupce (příznaky) datasetu. Tento dataset je dostupný na přiloženém CD pod názvem *df_2.csv*.

- *home_id, away_id* - Identifikátory domácího a hostujícího týmu.
- *home_league_rank, away_league_rank* - Umístění obou týmů v rámci celé soutěže.
- *home_conference_rank, away_conference_rank* - Umístění obou týmů v jejich konferencích.
- *home_division_rank, away_division_rank* - Umístění obou týmů v jejich divizích.
- *home_goals_scored, away_goals_scored* - Průměrný počet vstřelených branek obou dvou týmů na jeden zápas.
- *home_streak, away_streak* - Bilance posledních utkání. Např. pokud je *home_streak* roven 2, znamená to, že domácí mužstvo poslední dva odehrané zápasy vyhrálo. Naopak, pokud je *home_streak* roven např. -3, poslední tři odehrané zápasy domácí mužstvo prohrálo.
- *home_SAT, away_SAT* - Počet vstřel na bránu, mimo ní i zblokovaných (jedná se o údaj *corsi for*, popsaný v kapitole 1.2.1). Počet střel je vydělen počtem zápasů, takže finální údaj udává průměrný počet střel na jeden zápas.
- *home_USAT, away_USAT* - Počet vstřel na bránu, mimo ní, bez střel zblokovaných soupeřem (jedná se o údaj *fenwick*, popsaný v kapitole 1.2.2). Počet střel je vydělen počtem zápasů, takže finální údaj udává průměrný počet střel na jeden zápas.
- *home_saves, away_saves* - Počet úspěšných zásahů brankářů obou dvou týmů. Opět vydělen počtem zápasů, tudíž se jedná o průměrný počet úspěšných zákroků na jeden zápas.
- *home_goals_against, away_goals_against* - Počet inkasovaných branek obou dvou týmů. Opět se jedná o průměrný počet inkasovaných branek na jeden zápas.
- *home_PDO, away_PDO* - Moderní statistika popsaná v kapitole 1.2.3.
- *home_PIM, away_PIM* - Průměrný počet trestných minut každého týmu na jeden zápas.
- *home_corsi, away_corsi* - *Corsi* statistika obou dvou týmů, popsaná v kapitole 1.2.1. Také se jedná o průměrnou hodnotu na jeden zápas.
- *home_corsi_perc, away_corsi_perc* - Moderní statistika *Corsi percentage*, podrobně popsaná v kapitole 1.2.1.

- *home_fenwick, away_fenwick* - Průměrná hodnota statistiky *Fenwick*, popsané v kapitole 1.2.2.
- *home_shots, away_shots* - Průměrný počet střel na branku obou dvou týmů v jednom zápase.
- *home_games_played, away_games_played* - Počet odehraných zápasů v aktuální sezóně obou dvou týmů.
- *home_goalie, away_goalie* - Úspěšnosti zákroků obou dvou brankářů jednotlivých týmů.
- *home_goals_diff, away_goals_diff* - Průměrná hodnota, o kolik více či méně daný tým vstřelil branek, než inkasoval.
- *home_last_game, away_last_game* - Počet hodin od posledního odehraného zápasu.
- *home_odds, even_odds, away_odds* - Sázkové kurzy na výhru domácího týmu, na remízu a na výhru hostujícího týmu. Zdroj kurzů: `livesport.cz`
- *home_PO_chance, away_PO_chance* - Šanco obou dvou týmů na postup do playoff (vyřazovací části).
- *home_points, away_points* - Počet bodů v tabulce obou dvou týmů.
- *home_{X}_player, away_{X}_player* - Jedná se o 15 příznaků pro každý z týmů, celkem tedy 30 sloupců. {X} značí číslo od 0 do 14. V těchto sloupcích jsou hodnoty statistiky *Corsi percentage* patnácti nejlepších hráčů (domácího i hostujícího týmu), registrovaných pro utkání.
- *home_prev{X}, away_prev{X}* - Zde je {X} celé číslo od 1 do 5. Těchto celkem 10 sloupců značí, jak týmy hrály v předchozích pěti zápasech. Hodnoty ve sloupcích jsou pouze 0 v případě prohry a 1 v případě výhry. Např. home_prev_1 s hodnotou 1 značí, že poslední zápas domácí tým vyhrál a home_prev_2 s hodnotou 0 značí, že předposlední zápas domácí tým prohrál, a tak dále.
- *res* - Posledním sloupcem je sloupec s výsledkem utkání; 1 v případě výhry domácího týmu, 0 v případě výhry hostujícího týmu.

Kapitola 5

Experimenty

Tato kapitola se zabývá analýzou existujících řešení, experimenty nad datasetem, predikcemi výsledků různými modely a úspěšnostmi predikcí.

5.1 Existující řešení

Podobnému tématu, jako je téma této práce, se věnoval Pavel Suda ze Západočeské univerzity v Plzni, a to v bakalářské práci *Predikce vítěze sportovního utkání využitím PageRanku*. V této práci se autor zabýval využitelností algoritmu PageRank pro predikci (nejenom) výsledků zápasů v různých sportech. Autor se také věnoval predikci přímo Americké NHL. Autor při predikci použil jako vysvětlovanou proměnnou výsledek zápasu ve smyslu výhra/prohra/remíza. Tím se zmíněná práce liší od této práce, kde se zaměřuji převážně na predikce výsledku „do rozhodnutí", tedy vysvětlovaná proměnná nabývá pouze 2 hodnot - výhra a prohra. Zmíněná práce dosáhla v NHL úspěšnosti predikce 43,51%. [11]

Dalším, kdo se zabýval podobným tématem byl Filip Šimsa z Matematicko-fyzikální fakulty Univerzity Karlovy. Jeho diplomová práce *Analysis and prediction of league games results* se zabývá tvorbou predikčního modelu k identifikaci ziskových sázkových příležitostí. Autor se v této práci věnuje české nejvyšší hokejové lize, tedy Extralize. Jako vysvětlované proměnná je v této práci použit gólový rozdíl jednotlivých zápasů a jako model lineární regrese. [12]

Podobnému tématu se věnuje i práce Gianni Pischedda. Ten v roce 2014 publikoval práci *Predicting NHL Match Outcomes with ML Models*. V této práci se autor věnoval predikcím výsledků hokejových zápasů pomocí tří technik, a to rozhodovacích stromů, umělých neuronových sítí a softwaru ClusteR. Nejlepší úspěšnosti dosáhl software ClusteR, a to 61,54%. ClusteR je údajně software vyvinutý sázkovou kanceláří k predikci výsledků různých sportů a určením vhodných kurzů. Jediná informace k tomuto softwaru je, že kombinuje

několik technik strojového učení a jednou z nich je KNN (k-nearest neighbors, česky k-nejbližších sousedů). [13]

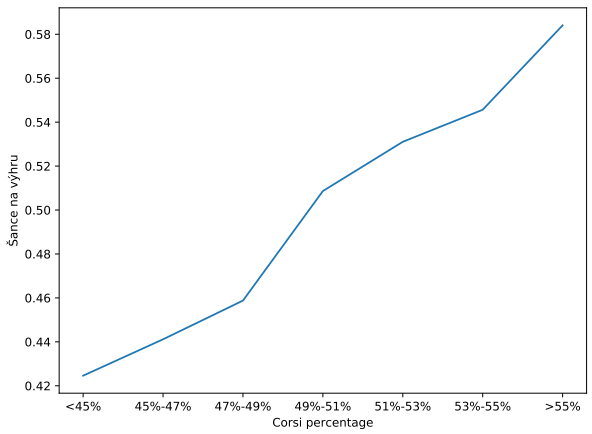
Poslední prací, kterou zmíním, je práce Martina Matuše z Vysoké školy ekonomické v Praze. Jedná se o práci *Predikce výsledků hokejových utkání pomocí data mining modelu*. Autor práce použil pro predikci neuronovou síť vytvořenou v aplikaci RapidMiner. Poté se pokusil předpovědět výsledky u 42 zápasů z Mistrovství světa 2014. Na těchto 42 zápasech měl autor přesnost predikce 71,43%. [14] Taková hodnota přesnosti predikce je velice dobrá, ale je však nutno říci, že se jedná o velmi malou sadu testovacích dat a také, že na Mistrovství světa jsou týmy méně vyrovnané, než v NHL. To lze vidět i na sázkových kurzech. Například při zápase Kanady a Jižní Koreje byl kurz na výhru Kanady 1,01 a na výhru Jižní Koreje 45. Takové rozdíly v NHL zpravidla nenastávají.

5.2 Vliv moderních statistik na výsledek utkání

Tato práce je zaměřena na získávání moderních hokejových statistik. V této části bude ukázána souvislost těchto statistik se samotným výsledkem utkání.

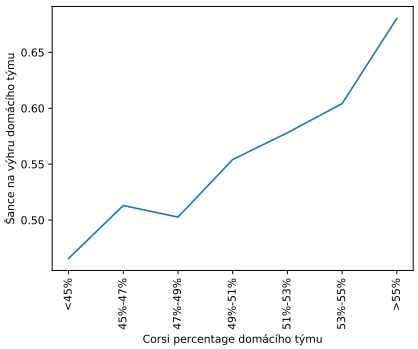
5.2.1 Corsi

První statistikou, jíž se práce zabývá, je statistika *Corsi* (potažmo *Corsi percentage*). Z dat, které byly v práci použity vyplývá, že s rostoucí hodnotou parametru *Corsi percentage*, roste i šance na výhru daného týmu. Při vynechání prvních 16 zápasů dané sezóny (první zápasy sezóny jsou vynechány kvůli obměně týmů po draftu a nemožnosti použít předchozí zápasy k získání důležitých příznaků a statistik, stejně tak jsou vynechány v predikčních modelech), je pravděpodobnost výhry týmu zobrazena na grafu 5.1.

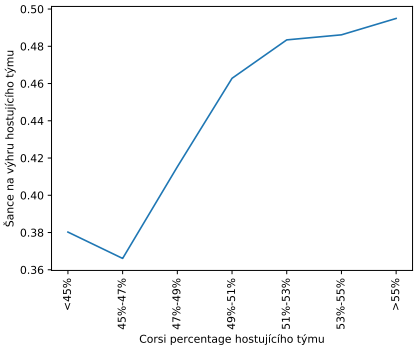


Obrázek 5.1: Pravděpodobnost výhry týmu v závislosti na jeho *Corsi percentage*.

Na grafu 5.1 vidíme, že pravděpodobnost výhry se přibližně pohybuje od 42 % do 58 %. Když se ovšem podíváme na následující graf 5.2, kde se berou v potaz pouze pravděpodobnosti výhry domácího týmu, jsou hodnoty výrazně vyšší. Hodnoty pravděpodobnosti se zde pohybují v rozmezí 45 % až do necelých 70 %. Naopak na grafu 5.3, kde se berou v potaz pouze pravděpodobnosti výhry hostujícího týmu vidíme, že jsou hodnoty pravděpodobnosti výrazně nižší, mezi 35 % a 50 %. Z těchto dat vyplývá, že vysoká hodnota *Corsi percentage* domácího týmu, nám může velmi pomoci při predikci výsledku utkání, převážně při predikci výhry domácího týmu.

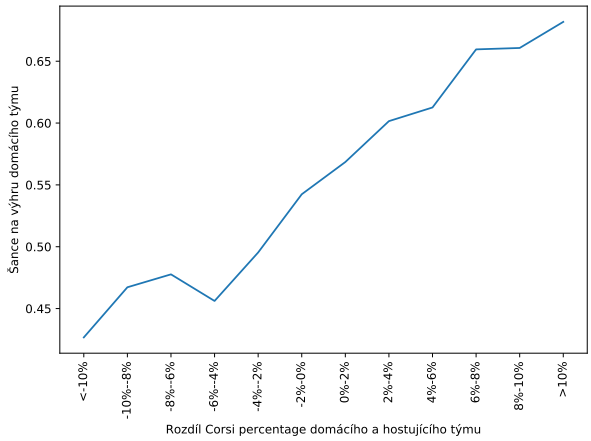


Obrázek 5.2: Pravděpodobnost výhry domácího týmu v závislosti na jeho *Corsi percentage*.



Obrázek 5.3: Pravděpodobnost výhry hostujícího týmu v závislosti na jeho *Corsi percentage*.

Na předchozích grafech jsme mohli vidět, jakou souvislost má *Corsi percentage* jednotlivých týmů a celkový výsledek zápasu. Na následujícím grafu 5.4, je zobrazen vliv rozdílu *Corsi percentage* domácího a hostujícího týmu na pravděpodobnost výhry domácího týmu. Z grafu 5.4 vyplývá, že pokud je rozdíl *Corsi percentage* větší než 10 % (ve prospěch domácího týmu), pravděpodobnost výhry domácích dosahuje téměř 70 %.

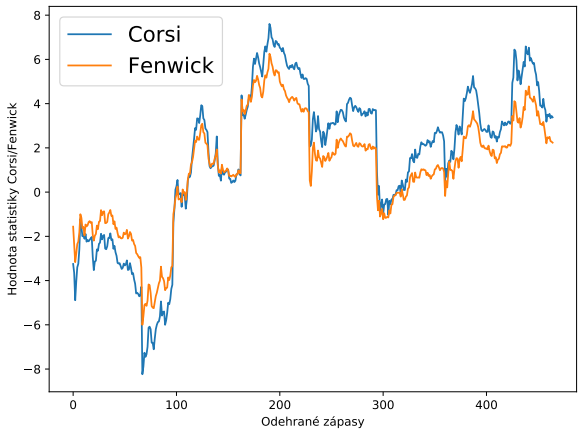


Obrázek 5.4: Pravděpodobnost výhry domácího týmu v závislosti na rozdílu *Corsi percentage*.

K těmto grafům je nutné dodat, že pravděpodobnosti výhry nebyly počítány za pomoci nějakého modelu. Jedná se o data z proběhlých utkání od začátku sezóny 2011/2012 až do 11. 2. 2019.

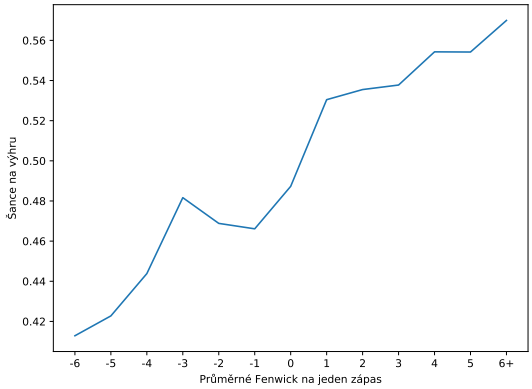
5.2.2 Fenwick

V případě statistiky *Fenwick*, byla předpokládána podobná závislost jako u statistiky *Corsi percentage*. To hlavně z důvodu, že jsou si obě statistiky velmi podobné. Podobnost statistik lze vidět i na následujícím grafu 5.5. V tomto grafu je zobrazen vývoj hodnot *Corsi* a *Fenwick* týmu *Tampa Bay Lighting* (dále jen *Tampa*). Na grafu 5.5 je vidět, že v prvních zápasech (sezóny 2011/2012 a 2012/2013) měla Tampa vyšší hodnotu *Fenwick*, než hodnotu *Corsi*. To znamená, že *Tampa* měla průměrně méně zblokovaných střel, než měl soupeř, což může značit to, že *Tampa* byla pod větším „tlakem". Obecně nižší hodnoty na začátku grafu 5.5 by měli značit neúspěch *Tampy*. To se potvrzuje i na celkových výsledcích soutěže z těchto sezón, kdy se *Tampa* umístila na 11. a 14. místě ve své konferenci, tedy ani v jedné ze zmíněných sezón *Tampa* nedosáhla ani na playoff. V následujících sezónách (kromě sezóny 2016/2017) se *Tampa* vždy playoff zúčastnila. Na grafu tedy můžeme vidět, že hodnoty *Corsi* a *Fenwick* opravdu korespondují se skutečnou formou/úspěšností týmu.



Obrázek 5.5: Vývoj hodnot statistik *Corsi* a *Fenwick* u týmu *Tampa Bay Lighting* od sezóny 2011/2012 do 11. 2. 2019.

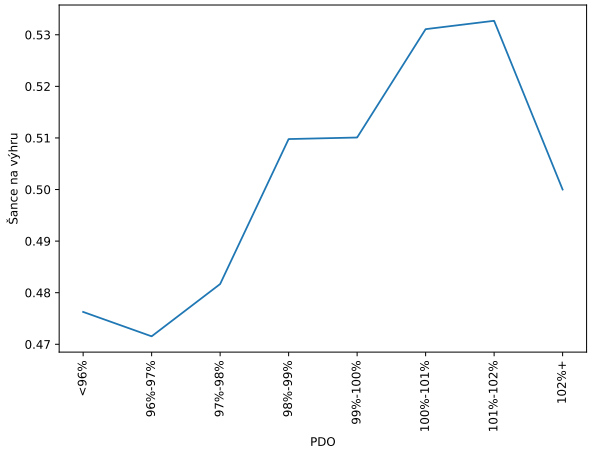
Na grafu 5.6 může čtenář opět vidět vývoj pravděpodobnosti výhry. V tomto grafu byly ovšem použity hodnoty statistiky *Fenwick* (na ose X), které byly vydělené počtem odehraných zápasů týmu. Výsledné hodnoty jsou tedy průměrné body *Fenwick* týmu na jeden zápas. Tyto hodnoty byly zaokrouhleny na celá čísla. V závislosti na nich je zobrazena pravděpodobnost výhry týmu.



Obrázek 5.6: Pravděpodobnost výhry týmu v závislosti na průměrném *Fenwick* na jeden zápas.

5.2.3 PDO

PDO je zajímavá statistika hlavně v tom ohledu, že by se nemělo jednat o statistiku „čím více, tím lépe“. Jak bylo zmíněno v části 1.2.3, pro statistiku *PDO* se používá i název *statistika štěstí* či *index štěstí*. To znamená že týmy s vyšší hodnotou této statistiky by neměli mít větší šanci na výhru. To částečně potvrzují i naše data. Na grafu 5.7 můžeme vidět, že nejvyšší pravděpodobnost úspěchu mají týmy s *PDO* mezi 100 % a 102 %. Týmům s vyšším *PDO* už ovšem šance na výhru klesají.



Obrázek 5.7: Pravděpodobnost výhry týmu v závislosti na jeho *PDO*.

5.3 Klasifikační modely

V této práci bylo využito několik klasifikačních modelů. Více modelů bylo použito, aby mohly být výsledky z těchto modelů porovnány nebo následně použity jako vstup modelu jiného. U všech modelů, s výjimkou *XGBoost*, byla použita jejich implementace v knihovně *scikit-learn*. Dále byly z této knihovny použity metody na předzpracování dat a výběr příznaků.

V práci byly použity základní modely, jako jsou *rozhodovací stromy, algoritmus KNN, logistická regrese* a *naivní Bayesův klasifikátor* i modely využívající kombinace více jednoduchých modelů, a to *rozhodovací lesy a XGB klasifikátor*. Při experimentech byl použit i model umělé neuronové sítě *MLPClassifier* (Multi-layer Perceptron classifier, klasifikátor vícevrstvý perceptron), který ovšem nedosahoval dobrých výsledků (pravděpodobně kvůli nedostatku dat), proto se jím práce více nezabývá.

Z datasetu popsaného v části 4.4, byly vytvořeny tři množiny dat (utkání). Jedná se o množinu trénovacích, validačních a testovacích dat. Tyto množiny

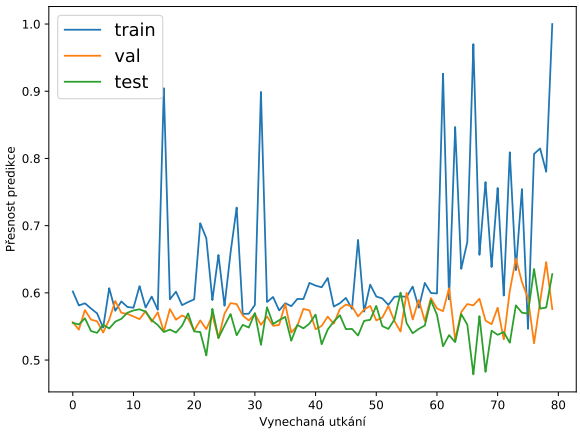
byly vytvořeny v poměru 56 : 19 : 25 (trénovací : validační : testovací). Množiny byly vytvořeny náhodně, pomocí funkce *train_test_split* z knihovny *scikit-learn, model_selection*.

5.3.1 Rozhodovací stromy

Pro práci s modelem rozhodovacího stromu byla použita implementace z knihovny *scikit-learn, DecisionTreeClassifier*. Jako hyperparametry modelu byly zvoleny následující proměnné.

- *max_depth* - Maximální hloubka naučeného stromu.
- *criterion* - Funkce měřící kvalitu rozdělení dat.
- *min_samples_split* - Minimální množství záznamů k provedení rozdělení dat.

Prvním experimentem bylo, kolik zápasů ze začátku sezóny je vhodné vynechat. Vynechání utkání ze začátku sezón bylo prováděno z důvodu, že po tzv. draftu, který probíhá po skončení soutěže, jsou týmy velmi odlišné než v sezóně předchozí. Mnoho hráčů tým opustí a naopak mnoho nových hráčů tým získá. Může tedy trvat několik utkání, než se tým „sehraje“ a získané hodnoty statistik se ustálí. Na grafu 5.8 je tedy zobrazena úspěšnost predikce *rozhodovacích stromů* v závislosti na počtu vynechaných prvních utkání v sezóně.



Obrázek 5.8: Vývoj přesnosti predikce v závislosti na počtu vynechaných prvních zápasů sezóny u *rozhodovacího stromu*.

V grafu 5.8 vidíme tři křivky. Modrá křivka je přesnost modelu na testovacích datech, oranžová na validační množině dat a zelená na testovacích

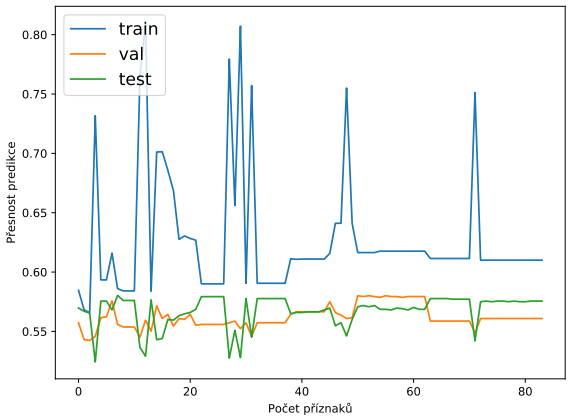
datech. Z grafu vyplývá, že model je nejpřesnější a nejstabilnější při vynechání 9 až 12 zápasů. Poté už jednotlivé křivky výrazně kolísají a navzájem se vzdalují. V pravé části grafu (při vynechání velkého počtu prvních utkání) lze vidět, že se přesnost na trénovacích datech blíží 80 % až 100 %. To značí tzv. přeučení modelu, které je způsobené malým množstvím dat v trénovacím datasetu a přílišnému přizpůsobení stromu datům v trénovacím datasetu.

Dalším experimentem provedeným nad daty a modelem byla standardizace dat. Použit byl tzv. StandardScaler z knihovny *scikit-learn, preprocessing.StandardScaler*, který data (jednotlivé příznaky) upraví tak, aby průměr hodnot byl 0 a směrodatná odchylka byla rovna 1. Tato standardizace pomáhá převážně u modelů, které nějakých způsobem poměřují „vzdálenost“ mezi jednotlivými datovými body (takový model je například *algoritmus KNN*). V našem případě jsou data částečně standardizovaná a to vydělením jednotlivých příznaků počtem odehraných utkání. Standardizace dat u tohoto modelu opravdu nepomohla, jak lze vidět v tabulce 5.1.

Standardizace	Trénovací	Validační	Testovací
Ne	61 %	56,08 %	57,55 %
Ano	61 %	56,08 %	57,50 %

Tabulka 5.1: Vliv standardizace příznaků na přesnost predikce *rozhodovacího stromu.*

V posledním experimentu s tímto modelem byl použit algoritmus *SelectKBest*. Jedná se o algoritmus, který ze všech dostupných příznaků vybere k příznaků, které by mohly mít největší vliv na vysvětlovanou proměnnou, přičemž k je hodnota zadaná uživatelem. Jedná se algoritmus z knihovny *scikit-learn, feature_selection.SelectKBest*. Na grafu 5.9 lze vidět, že nejvhodnější počet příznaků je 50 a více. Při nižších počtech příznaků sice občas model dosahoval větší přesnosti, ale často se model přeučil a výsledky nebyly stabilní a replikovatelné.



Obrázek 5.9: Vývoj přesnosti predikce v závislosti na počtu použitých příznaků z datasetu u *rozhodovacího stromu*.

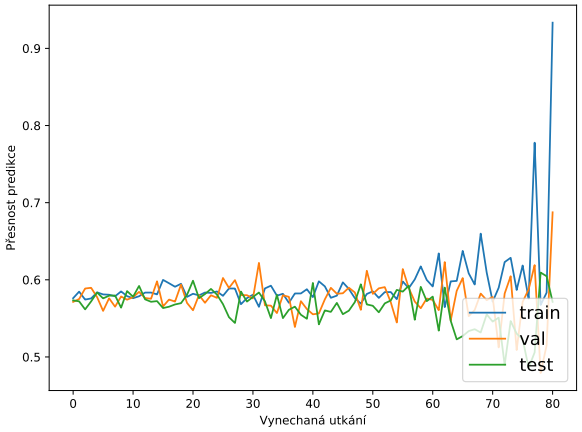
5.3.2 Logistická regrese

Jako další metoda pro predikci byla zvolena *logistická regrese*. Ačkoliv se metoda jmenuje regrese, ve skutečnosti se jedná o klasifikační metodu. Pro práci s modelem *logistické regrese* byla zvolena implementace z knihovny *scikit-learn, linear_model.LogisticRegression*. Jako hyperparametry modelu byly zvoleny následující proměnné.

- *fit_intercept* - Zda se má použít tzv. intercept.
- *C* - Míra regularizace (slouží ke snížení rizika přeučení modelu).

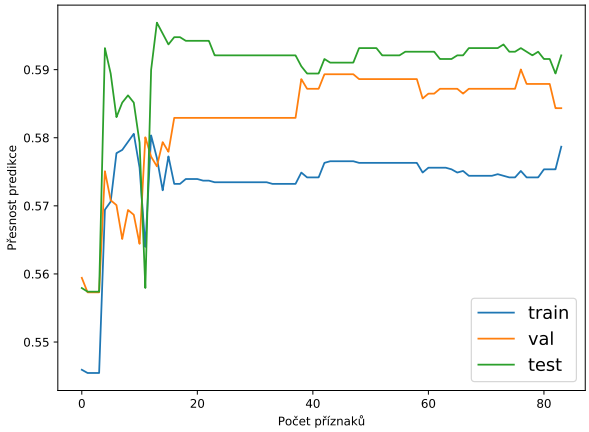
Při experimentu s vynechanými prvními zápasy sezóny, se model chová podobně jako model *rozhodovacího stromu*. Na grafu 5.10 je ovšem vidět, že model *logistické regrese* má výrazně nižší tendence se přeučit, než model *rozhodovacího stromu*. Také lze vidět o něco vyšší přesnost predikce, než u *rozhodovacího stromu*. Průměrná hodnota přesnosti na testovacích datech

u *rozhodovacího stromu* při vynechaných 5 až 15 prvních zápasů sezóny byla 56,21 % zatímco na stejných datech u *logistické regrese* se jednalo o hodnotu 57,73 %.



Obrázek 5.10: Vývoj přesnosti predikce v závislosti na počtu vynechaných prvních zápasů sezóny u *logistické regrese*.

U experimentu s algoritmem *SelectKBest* se model chová poněkud překvapivě. Jak lze vidět na grafu 5.11, tak v podstatě při jakémkoliv počtu vybraných příznaků, je přesnost na testovacím datasetu vyšší, než na trénovacím a validačním datasetu. To může být způsobeno nevhodným rozdělením dat na trénovací, validační a testovací množinu. Avšak používá se stejné rozdělení dat jako u ostatních modelů, kde tento jev pozorován nebyl.



Obrázek 5.11: Vývoj přesnosti predikce v závislosti na počtu použitých příznaků z datasetu u *logistické regrese*.

Co se týká standardizace příznaků, tak stejně jako u modelu *rozhodovacího stromu*, ani tady standardizace nepomohla. Standardizace příznaků u tohoto modelu vedla k lehkému zvýšení přesnosti na validační množině, ale k poklesu přesnosti na trénovací i testovací množině, jak lze vidět v tabulce 5.2.

Standardizace	Trénovací	Validační	Testovací
Ne	57,86 %	58,43 %	59,20 %
Ano	56,87 %	58,50 %	58,08 %

Tabulka 5.2: Vliv standardizace příznaků na přesnost predikce u *logistické regrese*.

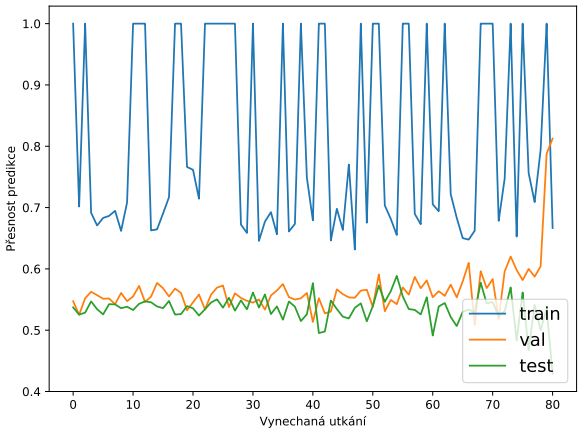
5.3.3 Algoritmus kNN

Model využívající algoritmus k-nejbližších sousedů, se od předchozích modelů liší. Zatímco u předchozích modelů trvalo dlouho samotné učení modelů a samotná klasifikace už byla rychlá, zde je tomu naopak. Algoritmus převádí jednotlivé záznamy (řádky datasetu) na n-dimenzionální vektory (kde n je počet příznaků) a poté počítá vzdálenost jednotlivých vektorů a vybere se k nejbližších sousedů (z testovací množiny), kteří poté „hlasují“ o výsledku klasifikace. Tudíž poté pro každý záznam v testovací množině je nutné spočítat, vzdálenost všech ostatních záznamů z množiny trénovací a vybrat z nich k nejbližších.

Pro práci s modelem využívajícím *algoritmus kNN* byla využita implementace z knihovny *scikit-learn, neighbors.KNeighborsClassifier*. Pro testování přesnosti predikce u tohoto modelu byly zvoleny (laděny) následující hyperparametry.

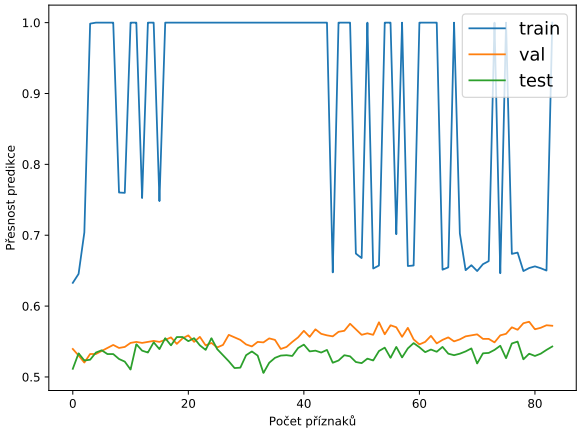
- *n_neighbors* - Počet nejbližších sousedů.
- *weights* - Váhová funkce, může např. upřednostňovat bližší sousedy při hlasování o finální klasifikaci.

Při experimentování s vynecháním prvních zápasů sezóny je jasně na grafu 5.12 vidět, že je model velmi náchylný k přeučení. Velice často se přesnost na trénovací množině dostala až ke 100 %, zatímco na validační a testovací množině se přesnost pohybovala pod 55 %. I nízká přesnost na testovacím datasetu ukazuje, že tento model není příliš vhodný pro naši klasifikační úlohu.



Obrázek 5.12: Vývoj přesnosti predikce v závislosti na počtu vynechaných prvních zápasů sezóny u *algoritmu kNN*.

Velmi podobně, jako na grafu 5.12, se chová přesnost predikcí i při použití algoritmu *SelectKBest*. I zde je vidět časté přeučení a přesnosti predikcí na testovací množině se pohybují mezi 50 % a 55 %. Vývoj přesnosti predikce v závislosti na počtu vybraných „nejlepších“ příznaků lze vidět a grafu 5.13.



Obrázek 5.13: Vývoj přesnosti predikce v závislosti na počtu použitých příznaků z datasetu u *algoritmu kNN*.

Obecně platí, že tento model je velmi náchylný na standardizaci příznaků. Proto zde bylo očekáváno výrazné zvýšení přesnosti predikce při použití standardizace příznaků. Ke zlepšení opravdu došlo, jak je vidět v tabulce 5.3.

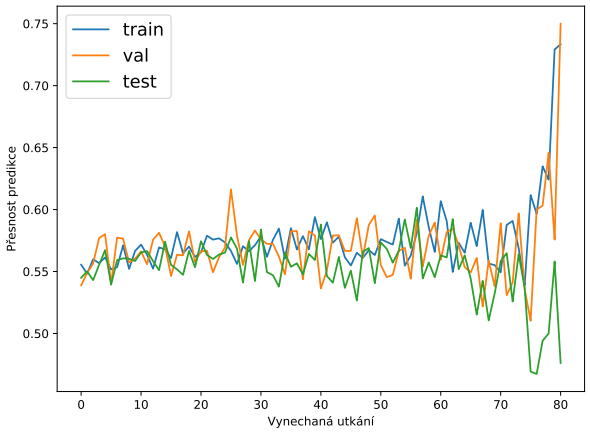
Standardizace	Trénovací	Validační	Testovací
Ne	100 %	53,59 %	53,28 %
Ano	100 %	57,22 %	54,30 %

Tabulka 5.3: Vliv standardizace příznaků na přesnost predikce u *algoritmu kNN*.

5.3.4 Naivní Bayesův klasifikátor

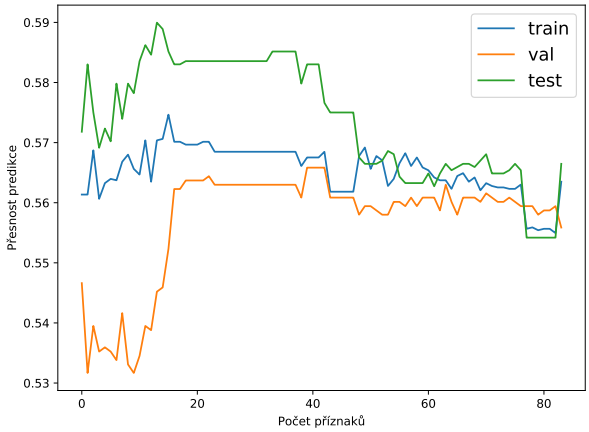
Posledním ze základních modelů je model *naivního Bayesova klasifikátoru*. Pro práci s modelem byla zvolena implementace z knihovny *scikit-learn, naive_bayes.GaussianNB*, ve kterém se počítá s *Gaussovo rozdělením* jednotlivých příznaků. Jako hyperparmatr zde byla zvolena jediná proměnná, a to *var_smoothing*.

Při experimentu s vynechanými prvním utkáními sezóny je vidět, že tento model má velice stabilní výsledky a není náchylný k přeučení. Na grafu 5.14 lze vidět, že přesnosti predikcí na všech množinách jsou si velmi blízko. Také lze pozorovat, že přesnost modelu na testovacím datasetu se pohybuje kolem 56%.



Obrázek 5.14: Vývoj přesnosti predikce v závislosti na počtu vynechaných prvních zápasů sezóny u *naivního Bayesova klasifikátoru*.

U experimentu se algoritmem *SelectKBest* se klasifikátor chová jinak, než u předchozího experimentu. Při použití méně než 50 „nejlepších“ příznaků, dosahuje přesnost modelu na testovací množině vyšších hodnot, než na validační a trénovací množině, jak můžeme vidět na grafu 5.15. To může být opět způsobeno nevhodným rozdělením dat. Při použití 50 a více příznaků, se hodnoty přesnosti predikcí na jednotlivých datasetech opět stabilizují a dávají podobné výsledky. Podobné chování bylo možno vidět i u *logistické regrese* na grafu 5.11.



Obrázek 5.15: Vývoj přesnosti predikce v závislosti na počtu použitých příznaků z datasetu u modelu *naivního Bayesova klasifikátoru*.

Standardizace příznaků má u tohoto modelu opět minimální vliv na přesnost predikce. Při standardizace příznaků se dokonce mírně sníží přesnost predikce, jak lze vidět v tabulce 5.4.

Standardizace	Trénovací	Validační	Testovací
Ne	56,35 %	55,59 %	56,65 %
Ano	56,06 %	55,59 %	56,43 %

Tabulka 5.4: Vliv standardizace příznaků na přesnost predikce u modelu *naivního Bayesova klasifikátoru*.

5.3.5 Náhodné lesy

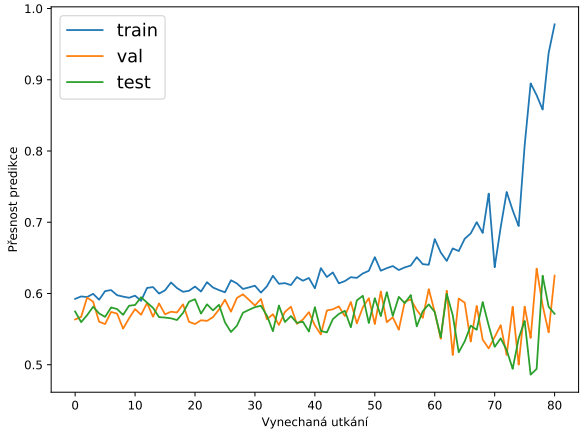
Dalším modelem, kterým se práce zabývá, je model *náhodného lesa*. Tento model se od předchozích modelů liší hlavně tím, že se jedná o tzv. *ensemble model*, tedy model, který kombinuje více jednoduchých modelů (v tomto případě *rozhodovacích stromů*). Implementace byla opět zvolena z knihovny *scikit-learn, ensemble.RandomForestClassifier*. Jako hyperparametry modelu byly zvoleny následující proměnné.

- *max_depth* - Maximální hloubka jednotlivých stromů lesa.
- *criterion* - Funkce měřící kvalitu rozdělení dat.
- *n_estimators* - Počet (rozhodovacích) stromů, které má daný les obsahovat.

Během experimentů s tímto modelem, se ukázalo, že pokud se nenastaví hodnota hyperparametru *max_depth* a nechá se na výchozí hodnotě, je model extrémně náchylný k přeučení. Jak je vidět na grafu 5.16, model dává celkem dobré a stabilní výsledky. Přesnost predikce na testovacím datasetu se pohybuje mezi 56% a 60%. Data pro grafy 5.16 a 5.17 byla získána s tímto nastavením hyperparametrů:

- *max_depth* = 3
- *criterion* = entropy
- *n_estimators* = 15, 25, 35, ..., 145

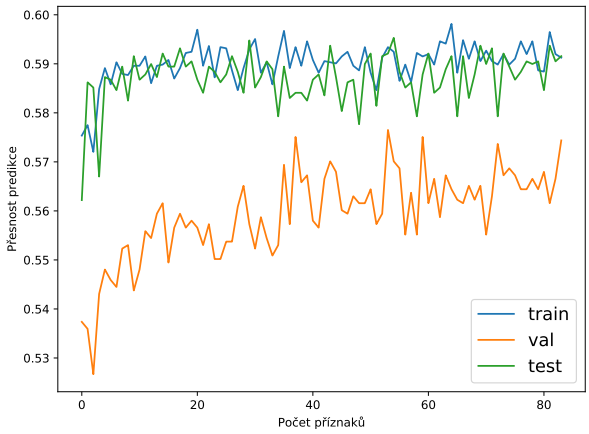
Pokud bychom nastavili nějaký rozsah i pro hodnotu *max_depth* a pro hyperparametr *criterion* bychom přidali hodnotu „gini“, čas potřebný k vytvoření grafu by exponenciálně vzrostl. Už s laděním tohoto jediného hyperparametru trvalo získat výsledky téměř 40 minut.



Obrázek 5.16: Vývoj přesnosti predikce v závislosti na počtu vynechaných prvních zápasů sezóny u *náhodných lesů*.

U tohoto modelu je nutné podotknout, že na dostupném hardwaru trvá samotné ladění hyperparametrů obecně velmi dlouho. V závislosti na zvoleném rozsahu jednotlivých hyperparametrů se jedná o jednotky až desítky minut.

U experimentu s algoritmem *SelectKBest* je na grafu 5.17 vidět, že se jedná o relativně stabilní model, který není náchylný na přeučení. Nižší přesnost u validačního datasetu může značit nevhodné rozdělení dat. Data ovšem pro všechny modely byla rozdělena stejně. Na grafu 5.17 je vidět, že nejvyšší přesnosti na trénovacím, testovacím a validačním datasetu dosahuje model kolem 25, 55 a 80 příznaků.



Obrázek 5.17: Vývoj přesnosti predikce v závislosti na počtu použitých příznaků z datasetu u *náhodných lesů*.

Standardizace dat v tomto případě mírně zvýšila přesnost predikce na testovacím datasetu. Na trénovacím a validačním datasetu se přesnost mírně snížila. Přesnost na testovacím datasetu se zvýšila o 0,32 %, jak je vidět v tabulce 5.5.

Standardizace	Trénovací	Validační	Testovací
Ne	78,21 %	57,79 %	58,78 %
Ano	64,78 %	57,44 %	59,10 %

Tabulka 5.5: Vliv standardizace příznaků na přesnost predikce u *náhodného lesa*.

5.3.6 XGBoost klasifikátor

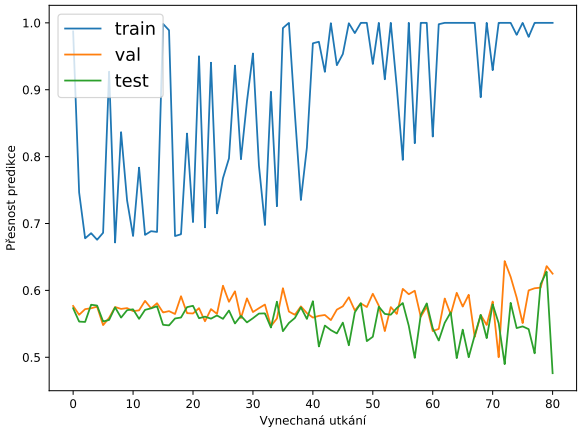
Posledním modelem, který bude v práci zmíněn je opět *ensemble model*. Jedná se model klasifikátoru *XGBoost* (Extreme Gradient Boosting), který pracuje s rozhodovacími stromy, podobně jako náhodné lesy. Tento model získal velkou

popularitu a je velmi oblíbený i mezi uživateli serveru `kaggle.com` (populární web zabývající se strojovým učením a „data science“), kde v několika soutěžích tento model zvítězil jako nejpřesnější. [15]

Při práci s tímto modelem bylo nutné počítat s časově náročnějším učením. Ze všech předchozích modelů zde byly časy pro naučení a otestování modelu nejdelší. Řádově se jednalo o desítky minut až jednotky hodin. Data pro každý z následujících grafů trvalo získat více než 2 hodiny. Jako hyperparametry tohoto modelu byly zvoleny pouze dvě následující proměnné. Při dřívějším experimentování bylo použito i více hyperparametrů, avšak čas potřebný k naučení modelu na těchto hyperparametrech se pohyboval v řádu několika hodin a výsledky nebyly o mnoho lepší, než při ladění pouze těchto dvou hyperparametrů.

- *max_depth* - Maximální hloubka jednotlivých stromů lesa.
- *n_estimators* - Maximální počet (rozhodovacích) stromů, které daný model naučí.

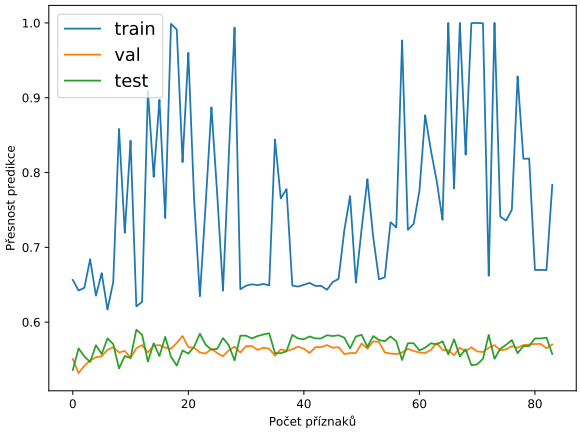
Při experimentu s vynechanými prvními zápasy sezóny se ukazuje, že nejvyšší přesnosti a stability dosahuje model při vynechání prvních 5 až 12 utkání. při tomto počtu utkání dosahuje model až 59 % přesnosti. Na grafu 5.18 lze vidět tendenci modelu k přeučování. Obecně zde přesnost na trénovacím datasetu neklesá pod 66 %. Podobné chování jsme pozorovali i u modelu využívající *algoritmus kNN*.



Obrázek 5.18: Vývoj přesnosti predikce v závislosti na počtu vynechaných prvních zápasů sezóny u *XGBoost klasifikátoru*.

U experimentu s algoritmem *SelectKBest* a modelu *XGBoost* lze vidět, že tento algoritmus výběru příznaků má na přesnost minimální vliv. Přesnost na

trénovacím datasetu se sice pohybuje od 60 % až do 100 %, avšak přesnosti na testovacím a validačním datasetu se pohybují pouze minimálně. Na grafu 5.19 lze vidět, že model dosahuje nejvyšší přesnosti na testovacím datasetu kolem 59 %.



Obrázek 5.19: Vývoj přesnosti predikce v závislosti na počtu použitých příznaků z datasetu u *XGBoost klasifikátoru*.

Při použití standardizace dat byly očekávány podobné výsledky jako u modelu *náhodného lesa*, neboť oba modely pracují s modely rozhodovacích stromů. U tohoto modelu nastala zajímavá situace. Vliv standardizace příznaků se ukázal jako úplně nulový. Při použití i nepoužití standardizace příznaků byla přesnost na trénovacím datasetu 61,57 %, na validačním 57,95 % a na testovacím 58,43 %.

5.4 Porovnání modelů

Z experimentů na různých modelech vyplývá, že nejvyšší přesnost na testovacím datasetu byla dosažena u modelu *náhodného lesa*. V tabulce 5.6, jsou zobrazeny přesnosti jednotlivých modelů s použitím metod předzpracování dat a výběru příznaků.

Z tabulky 5.6 vyplývá, že nejvyšší přesnost na testovacím datasetu byla 59,58 % a to u modelu *náhodného lesa* s použitím 79 nejlepších příznaků a se stanardizací dat. Podobně úspěšný byl i model *logistické regrese* s přesností 59,32 %, při použití 68 nejlepších příznaků a bez použití normalizace dat. Naopak nejnižší predikční schopnosti (s ohledem na přesnost) má model s použitím *algoritmu kNN* a to s přesností 53,28 %. Všechny modely byly učeny na stejně rozděleném datasetu a tudíž je možné jejich výsledky porovnávat.

Model	Standardizace	SelectKBest	Trénovací	Validační	Testovací
Náhodný les	Ano	79	59,46 %	57,86 %	59,58 %
Náhodný les	Ne	79	59,57 %	57,65 %	59,16 %
Logistická regrese	Ano	68	57,06 %	58,36 %	58,25 %
Logistická regrese	Ne	68	57,44 %	58,72 %	59,32 %
Rozhodovací strom	Ano	65	61,14 %	55,87 %	57,71 %
Rozhodovací strom	Ne	65	61,14 %	55,87 %	57,77 %
Algoritmus kNN	Ano	84	100 %	57,22 %	54,30 %
Algoritmus kNN	Ne	84	100 %	53,59 %	53,28 %
Naivní Bayes	Ano	76	56,06 %	55,59 %	56,43 %
Naivní Bayes	Ne	76	56,35 %	55,59 %	56,65 %
XGBoost	Nemá vliv	84	61,57 %	57,95 %	58,43 %

Tabulka 5.6: Porovnání přesnosti jednotlivých modelů

5.5 Kombinace výsledků modelů

Po provedení všech experimentů, byly výsledky predikcí jednotlivých modelů (kromě modelu vyžívající *algoritmus kNN*, který byl výrazně nejméně úspěšný) pro testovací dataset (1873 utkání) uloženy do souborů. Do těchto souborů byly uloženy pravděpodobnosti jednotlivých predikcí (tedy hodnoty 0 až 1). Tento dataset byl opět rozdělen na trénovací a testovací část. Na trénovací části datasetu byly zkoumány různé kombinace zmíněných pravděpodobností. Tyto pravděpodobnosti byly náhodně i ručně kombinovány a byl tak vytvořen model, jehož úspěšnost se pohybovala kolem 60 % (vzhledem k náhodnému výběru vah predikcí jednotlivých modelů se přesnost pohybovala od 59 % do 61 %). Tato kombinace výsledků předchozích modelů dává již velmi slibnou přesnost, která se blíží přesnosti predikcí sázkových kanceláří. Obvyklé hodnoty vah pro predikce jednotlivých modelů lze vidět v tabulce 5.7. (S těmito vahami byla finální přesnost predikce 59,4 % na trénovacím datasetu a 60,34 % na testovacím datasetu)

Model	Váha
Rozhodovací strom	<1 %
Logistická regrese	20 %
Naivní Bayes	<1 %
Náhodné lesy	62 %
XGBoost	17 %

Tabulka 5.7: Váhy predikcí jednotlivých modelů při tvorbě výsledného modelu

5.6 Porovnání predikcí s predikcemi sázkových kanceláří

Po zvolení nejlepšího modelu (modelu s nejvyšší přesností) výsledky ukazují, že přesnost na testovacím datasetu byla 59,58 %. Na stejném datasetu byla přesnost predikce sázkových kanceláří 60,76 %. Přesnost predikce sázkových kanceláří byla počítána pouze podle kurzů. Pokud byl sázkový kurz na výhru domácího týmu nižší, než na výhru hostů, byla za predikci sázkové považována výhra domácího mužstva a naopak, byl-li sázkový kurz na výhru hostů nižší, než na výhru domácích, byla za predikci považována výhra hostujícího mužstva. Nutno podotknout, že náš výsledný model také nedává vždy stejné výsledky (používá v průběhu nedeterministické algoritmy) a jeho přesnost se pohybuje přibližně od 58,5 % až do 61 %.

Dále se ukázalo, že náš model se se sázkovou kanceláří shodl v 83,88 % utkání, z toho obě predikce byly úspěšné v 62,12 % utkání a neúspěšné v 37,88 % utkání. Z toho také vyplývá, že pokud se náš model „shodne“ s predikcí sázkové kanceláře, je přesnost predikce kolem 63 %. Stále zde ale existuje několik (16,12 %) utkání, kde se modely neshodnou. V těchto utkáních je v 52,65 % utkání úspěšná predikce sázkových kanceláří a v 47,35 % utkání je úspěšná predikce našeho modelu. Zde je největší prostor pro zlepšení přesnosti predikce. Bylo by potřeba identifikovat ty zápasy, kde je přesnější model sázkové kanceláře a kde náš. Jedná se o utkání, která pravděpodobně lze správně predikovat (neboť je buď sázková kancelář, nebo náš model predikuje správně). Nalezením tohoto „klíče“, by se mohla přesnost predikce (v ideálním případě) zvýšit až na více než 70%. Další možností by bylo, zaměřit se na utkání, kde se náš model a model sázkové kanceláře neshodnou a v těchto případech identifikovat utkání, kde je přesnější náš model. To by vedlo k vytvoření modelu, který by dokázal určit utkání s potenciálem nejvyššího zisku.

Závěr

. V závěru práce lze říci, že existuje relativně široké spektrum zdrojů informací k proběhlým zápasům NHL. Nejobsáhlejší z těchto zdrojů je přímo web `nhl.com`, který i ostatní zdroje (většinou webové stránky) zpracovávají a upravený obsah opět poskytují. Existují i zdroje, které poskytují přímo moderní statistiky, jako jsou *Corsi, Fenwick* a *PDO*. Ovšem z těchto zdrojů je náročné data stáhnout, neboť nemají volně dostupné API. I z toho důvodu byl zvolen za hlavní zdroj web `nhl.com`, ze kterého lze data jednoduše stáhnout a poté z nich spočítat moderní statistiky.

Získání dostupných dat o proběhlých zápasech z webu `nhl.com`, se podařilo podle očekávání. Náročnější byl získání historických sázkových kurzů k proběhlým zápasům. Na konec pro tento účel posloužil web `livesport.cz`, kde však bylo nutné získat povolení ke stažení dat od provozovatele. Stažení těchto dat bylo komplikovanější i tím, že data nebyla dostupná pomocí nějakého API, ale bylo nutné napsat si vlastní parser, který informace potřebné ke stažení kurzů získá přímo ze zdrojového kódu stránek.

Po stažení dat a jejich zpracování bylo nutné projít jednotlivé příznaky a zaměřit na jejich vliv na konečný výsledek zápasu. V práci byl zjišťován vliv převážně moderních hokejových statistik, a to v kapitole 5.2. Bylo zjištěno, pomocí těchto statistik lze odhadovat výsledek zápasu a že existuje spojitost mezi hodnotou jednotlivých statistik a pravděpodobností výhry jednotlivých týmů.

V práci bylo použito několik klasifikačních modelů. Přesnost predikcí se u jednotlivých modelů lišila v rámci 5%. Jako nejstabilnější se ukázal model *náhodného lesa*, který trvale předpovídal výsledky s přesností mezi 58,5% a 61%.

V práci byl také zkoumán vliv použití moderních hokejových statistik na přesnost predikce. Ačkoliv se v části 5.2 ukázala souvislost mezi moderními statistikami a šancí na výhru, tak celkový vliv byl menší, než se očekávalo. Při vynechání příznaků obsahující moderní hokejové statistiky klesla přesnost predikce pouze minimálně. Nejvíce přesnost klesla při vynechání těchto příznaků

u modelu *logistické regrese* a to přibližně o 0,5 %. Nejmenší vliv měly moderní statistiky na model *náhodného lesa*, kde při vynechání těchto příznaků klesla přesnost predikce pouze o 0,1 %.

Dále byla v práci využita i vlastnost jednotlivých modelů, že v implementace modelů v knihovně *scikit-learn* umožňuje vracet pravděpodobnosti predikce. Tyto pravděpodobnosti byly uloženy do souborů a poté byl vytvořen model, který kombinoval výsledky jednotlivých modelů a tvořil finální predikci. Tato výsledná predikce měla úspěšnost přibližně 60%.

Přesnost predikce sázkových kanceláří byla 60,76 %. Této hodnotě se přiblížili i některé ze zkoumaných modelů. Jedná se o úspěšný experiment, neboť sázkové kanceláře by už ze své podstaty měly mít nejlepší predikce. Je zde i velký prostor pro zlepšení přesnosti teoreticky až k 70 %. Této přesnosti by bylo možné dosáhnout, pokud by byl nalezen „klíč“, který by určil, kdy je přesná predikce sázkové kanceláře a kdy predikce našeho modelu. To ovšem není jednoduchá úloha a těžko říct, jestli takový klíč existuje a je v silách člověka a strojového učení ho nalézt.

Práce, které se zabývají tématem predikce výsledků hokejových zápasů v NHL existují, ale není jich velké množství. Většina těchto prací s pohybuje s přesností predikce kolem 60%. Zde je prostor na tuto práci navázat. V NHL existuje velké množství zdrojů informací k jednotlivým utkání, týmům i hráčům. Pokud by se podařilo zpracovat tato data (např. z webových stránek jednotlivých týmů, z účtů hráčů na sociálních sítích, z různých internetových fór) a efektivně je využít, byl by to další a neméně podstatný zdroj informací k predikci výsledků zápasů. Co se NHL týče, existují práce, které se zabývají predikcí platu a pořadí draftu hráčů a k těmto pracím existuje i velké množství datasetů (většinou se jedná o souhrn statistik za proběhlé sezóny jednotlivých hráčů), ale bohužel tyto datasety jsou jen těžce využitelné pro predikci výsledků jednotlivých zápasů.

Bibliografie

1. MARSH, James H. *National Hockey League (NHL)* [online]. 2017 [cit. 2019-04-03]. Dostupné z: https://www.thecanadianencyclopedia.ca/en/article/national-hockey-league.
2. KLEIN, Cutler. *From six teams to 31: History of NHL expansion* [online]. 2016 [cit. 2019-04-03]. Dostupné z: https://www.nhl.com/news/nhl-expansion-history/c-281005106.
3. HORGAN, Colin. *NHL lockout: How it came to this and what happens next* [online]. 2012 [cit. 2019-04-05]. Dostupné z: https://www.theguardian.com/sport/blog/2012/sep/17/nhl-lockout-2012-season-delayed.
4. MJEX19. *THE IDEAL NHL PLAYOFF FORMAT* [online]. 2017 [cit. 2019-04-09]. Dostupné z: https://matthewjex.wordpress.com/2017/10/30/the-ideal-nhl-playoff-format/.
5. HAMMOND, Richard. *Understanding Advanced Stats, Part One: Corsi and Fenwick* [online]. 2011 [cit. 2019-04-03]. Dostupné z: https://www.matchsticksandgasoline.com/2011/7/29/2290643/understanding-advanced-stats-part-one-corsi-fenwick.
6. BURTCH, Steve. *Intro To Advanced Hockey Statistics - Fenwick* [online]. 2012 [cit. 2019-04-03]. Dostupné z: https://www.pensionplanpuppets.com/2012/7/25/3184137/intro-to-advanced-hockey-statistics-fenwick.
7. NINJA, Hockey. *Měříme "štěstíčko": Které týmy TELH nemají udržitelné výsledky?* [online]. 2012 [cit. 2019-04-04]. Dostupné z: https://www.hokej.cz/merime-stesticko-ktere-tymy-telh-nemaji-udrzitelne-vysledky/5012703.
8. POTHIER, Jason. *Advanced Stats For VGK Dummies: PDO* [online]. 2017 [cit. 2019-04-04]. Dostupné z: https://sinbin.vegas/advanced-stats-vgk-dummies-pdo/.

9. LIVESPORT.CZ. *Všeobecné podmínky užívání,7 Duševní vlastnictví* [online]. 2018 [cit. 2019-04-09]. Dostupné z: https://www.livesport.cz/podminky-uziti/.

10. HYNES, Drew. *NHL Stats API Documentation* [online]. 2018 [cit. 2019-04-05]. Dostupné z: https://gitlab.com/dword4/nhlapi/blob/master/stats-api.md.

11. SUDA, Pavel. *Predikce vítěze sportovního utkání využitím PageRanku* [online]. 2014 [cit. 2019-04-12]. Dostupné z: https://otik.zcu.cz/bitstream/11025/13540/1/BP_%20sudap_A11B0612P.pdf.

12. ŠIMSA, Filip. *Analysis and prediction of league games results* [online]. 2015 [cit. 2019-04-13]. Dostupné z: https://is.cuni.cz/webapps/zzp/detail/130787/.

13. PISCHEDDA, Gianni. *Predicting NHL Match Outcomes with ML Models* [online]. 2014 [cit. 2019-04-13]. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.735.795&rep=rep1&type=pdf.

14. MATUŠ, Martin. *Predikce výsledků hokejových utkání pomocí data mining modelu* [online]. 2015 [cit. 2019-04-13]. Dostupné z: https://vskp.vse.cz/44951_predikce_vysledku_hokejovych_utkani_pomoci_data_mining_modelu.

15. SYNCED. *Tree Boosting With XGBoost — Why Does XGBoost Win "Every" Machine Learning Competition?* 2017. Dostupné také z: https://medium.com/syncedreview/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition-ca8034c0b283.

Příloha A

Seznam použitých zkratek

API Application User Interface

IDE Integrated Development Enviroment

JSON JavaScript Object Notation

kNN k-Nearest Neighbors

NBA National Basketball Association

NFL National Football League

NHA National Hockey Association

NHL National Hockey League

NHLPA National Hockey League Player Association

REST Representational state transfer

URL Uniform Resource Locator

Příloha B

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `src`
  - `skript` ..... skripty použité v práci
  - `dataset` ..... finální dataset a predikce jednotlivých modelů
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
  - `souhlas.png` ..... souhlas s použitím dat z webu `livesport.cz`
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF
  - `thesis.ps` ..... text práce ve formátu PS