SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE
WARSAW SCHOOL OF ECONOMICS

Studium Licencjackie

Kierunek **MIESI**
Specjalność **Metody Analizy Decyzji**

Imię i nazwisko autora **Maciej Beręsewicz**
Nr albumu **72297**

Tytuł pracy

Przewidywanie wyników meczów piłkarskich w Bundeslidze za pomocą algorytmów uczenia maszynowego

Praca licencjacka
napisana w Katedrze/Instytucie
Ekonometrii

pod kierunkiem naukowym
dr hab. Michał Rubaszek, prof. SGH

Warszawa 2019 – 11 - 05

Spis treści

Wstęp

Ogromne ilości danych docierających codziennie ze wszystkich stron dotyczą również jednego z najpopularniejszych obecnie sportów świata, piłki nożnej. Niemal niemożliwe jest obejrzenie serwisu informacyjnego, skorzystanie z portalu społecznościowego, czy internetowych mediów bez otrzymania garści statystyk dotyczących tego sportu. Na stronach internetowych nie sposób przeoczyć wielkie nagłówki typu: „Dwa gole i asysta Roberta Lewandowskiego w meczu z Schalke"; albo: „Sergio Ramos rekordzistą w liczbie czerwonych kartek w historii La Liga".[1] Dostępne dane piłkarskie dotyczyć mogą niemal wszystkiego: od zwykłego wyniku meczu, po liczbę oddanych strzałów na bramkę, czy liczbę przebiegniętych kilometrów przez obrońców drużyny gości. Kiedy ma się chęć poznania dowolnej statystyki dotyczącej piłki nożnej, to istnieje bardzo duża szansa, że po przeszukaniu internetu odpowiedź na postawione pytanie będzie dostępna. O szerokiej dostępności danych piłkarskich świadczy również fakt, że komentatorzy piłkarscy podczas niemal każdego meczu są w stanie zasypać widza różnego typu ciekawostkami i zależnościami. Słyszy się przykładowo, że: „Borussia Dortmund nigdy w historii nie przegrała meczu Bundesligi rozgrywanego w piątek"; albo: „Bayern Monachium w każdym meczu w tym sezonie z drużynami w zielonych koszulkach strzelał co najmniej 2 bramki" [2]. Celem poniższej pracy będzie próba znalezienia zmiennych, które faktycznie mają wpływ na rezultat pojedynczego spotkania i budowa modelu przewidującego wyniki meczów piłkarskich. Z pewnością zmienne typu „dzień tygodnia rozgrywania meczu", czy zmienna „kolor koszulek" zostaną w modelu pominięte. Do zagadnienia podejdę z dwóch stron: z jednej strony model podejmie próbę przewidzenia rezultatu spotkania (wygrana gospodarza, remis lub wygrana gościa), co może zostać wykorzystane przez osobę chcącą typować wyniki meczów. Z drugiej strony model będzie również przewidywał prawdopodobieństwa powyższych zdarzeń i może zostać wykorzystany przez firmę bukmacherską.

W pierwszym rozdziale przyjrzę się bliżej lidze, na podstawie której model zostanie zbudowany i do przewidywania której będzie mógł zostać wykorzystany. Ligą tą jest niemiecka Bundesliga. Następnie opiszę wykorzystane przeze mnie dane dotyczące historycznych meczów Bundesligi oraz historycznych kursów bukmacherskich na te mecze i przekształcenia, jakich dokonałem w celu przygotowania zestawu gotowego do budowy modelu. Dodatkowo, na podstawie analiz, spróbuję znaleźć pewne zależności i wyliczyć statystyki, które pomogą później w budowie modelu i ocenie jego jakości.

Drugi rozdział dotyczyć będzie zastosowanych metod badawczych i budowy modelu. Opiszę wykorzystane do treningu modelu algorytmy lasu losowego oraz sieci neuronowych oraz porównam różnice w wynikach między tak zbudowanymi modelami.

W rozdziale trzecim porównam wyniki modelu uzyskanego za pomocą sieci neuronowych do wyników prostych modeli, które można zamiennie zastosować (takie jak: typowanie zawsze wygranej drużyny domowej) oraz modeli wykorzystywanych przez największe firmy bukmacherskie na świecie. Będą to dobre punkty odniesienia dla zbudowanego przeze mnie modelu.

Rozdział 1 – Opis zjawiska

Dlaczego Bundesliga?

Standardowe pytanie, jakie często można otrzymać, kiedy rozmówca dowiaduje się, że interesujesz się piłką nożną, brzmi: ”Jesteś za Realem, czy Barceloną?”. I rzeczywiście, zgodnie z rankingiem[3] opublikowanym w sierpniu 2018 roku na portalu „Businessinsider”, te dwa Hiszpańskie kluby mają zdecydowanie największą liczbę kibiców mierzoną ilością followersów na portalach społecznościowych. Zadający to pytanie być może nie rozumieją jednak, że istnieją również inne ligi piłkarskie i ligi te mogą być równie ciekawe, jak liga hiszpańska (będąca co trzeba uczciwie przyznać najlepszą ligą na świecie, konkurować z nią może tylko angielska Premier League). Ja natomiast od siedmiu lat jestem kibicem Bundesligi, którą uważam za najciekawszą ligę świata. To prawda, zespoły niemieckie w ostatnich latach w europejskich pucharach regularnie odpadają na wcześniejszych etapach rozgrywek, niż drużyny hiszpańskie, angielskie, czy włoskie. To prawda, że liga może wydawać się monotonna, gdyż od kilku lat mistrzem zostaje zawsze Bayern Monachium (co dla mnie jako kibica Borussii Dortmund jest szczególnie bolesne). Niemiecka piłka jest jednak szybka, ofensywna i, wyłączając mecze Bayernu Monachium, względnie nieprzewidywalna. Każdy zespół w Bundeslidze ma szansę odnieść zwycięstwo z każdym innym. W lidze hiszpańskiej natomiast co roku te same trzy zespoły zajmują trzy pierwsze miejsca (Barcelona, Real i Atletico). W lidze angielskiej zazwyczaj te same sześć zespołów zajmuje sześć pierwszych miejsc (wyjątkiem mistrzowski sezon 2015/16 w wykonaniu Leicester City). W związku z tym w europejskich pucharach grają co roku niemal te same drużyny. W Bundeslidze natomiast jedynie Bayern Monachium i Borussia Dortmund regularnie grają w pucharach. Liga niemiecka wydaje się zatem względnie nieprzewidywalna, co czyni jej modelowanie dużym wyzwaniem.

Utworzony przeze mnie model bazuje na danych dotyczących meczów Bundesligi od roku 1993 aż do roku obecnego. Bundesliga od sezonu 1993/94 ma tę szczególną i ułatwiającą modelowanie cechę, że jest stosunkowo stabilna. Mam tu na myśli:

1. Fakt, że właśnie od sezonu 1993/94 aż do chwili obecnej za zwycięstwo w meczu Bundesligi przyznawane są 3 punkty, za remis 1 punkt, za porażkę 0 punktów. Przed sezonem 1993/1994 za zwycięstwo przyznawane były tylko 2 punkty. Wykorzystanie danych sprzed tego sezonu byłoby zatem mocno utrudnione i istniałaby konieczność pewnej standaryzacji liczby punktów.
2. Fakt, że liczba drużyn grających w Bundeslidze jest co sezon taka sama (18 drużyn, do sezonu 1991/92 było to 20 drużyn) i zawsze rozgrywana jest taka sama liczba meczów

(inaczej niż na przykład w polskiej Ekstraklasie, gdzie od kilku sezonów niektóre drużyny grają ze sobą dodatkowy mecz zależnie od tego, czy w rundzie zasadniczej skończyły sezon w górnej, czy w dolnej połowie tabeli).

3. Fakt, że stosunkowo niewielka liczba meczów została przełożona z powodu warunków pogodowych oraz innych czynników.

Powyższe cechy bardzo ułatwiają modelowanie Bundesligi. Liga niemiecka ma jednak również cechy utrudniające jej modelowanie. Są to:

1. Wspomniana wcześniej względna nieprzewidywalność wyników meczów.
2. Fakt, że w niektórych sezonach do Bundesligi awansowały trzy drużyny a w niektórych tylko dwie w zależności od wyniku meczu barażowego pomiędzy szesnastą drużyną Bundesligi a trzecią drużyną 2. Bundesligi.

Podczas budowy modelu każda zmienność utrudnia dokonanie właściwej predykcji, zatem cechy te nieco komplikują analizę. Ogólnie jednak, liga niemiecka wydaje się bardziej stabilna niż niektóre inne ligi, na przykład liga hiszpańska (znacznie więcej przełożonych meczów, liczba drużyn w niektórych sezonach inna niż w pozostałych). Natomiast głównym powodem wybrania Bundesligi była chęć jej głębszego poznania i zrozumienia.

Opis danych i przekształceń dokonanych na danych

Opisanie wszystkich przekształceń i uproszczeń jakie zostały dokonane na danych zajęłoby bardzo dużo czasu. Opiszę zatem tylko po krótce postać zbioru danych, sposób utworzenia poszczególnych zmiennych oraz metody zastosowane do uzupełnienia braków danych.

Wyjściowe zbiory danych zostały pobrane z bazy danych Football-Data[4]. Każdy zbiór dotyczy pojedynczego sezonu Bundesligi od sezonu 1993/94 do sezonu 2018/19 i składa się z 22 zmiennych. Dane z sezonów wcześniejszych niż sezon 1993/94 nie zostały przeze mnie wykorzystane ze względu na cechy opisane w podrozdziale „Dlaczego Bundesliga?”. Dane składają się ze zmiennych wskazujących drużynę domową i drużynę wyjazdową oraz datę rozgrywania meczu, jak również rezultat oraz liczbę bramek strzelonych przez obie drużyny. Dodatkowo dołączone zostały historyczne kursy kilkunastu największych światowych bukmacherów dotyczące każdego z meczów.

Powyższe dane zostały wykorzystane jako baza do utworzenia zmiennych użytecznych w prognozowaniu wyników meczów. Zmienne utworzone przeze mnie bazują wyłącznie na liczbie punktów zdobytych przez daną drużynę w danym historycznym meczu. Dane mają charakter szeregu czasowego, gdyż każda kolejna obserwacja zależy od pewnych obserwacji

poprzedzających je w czasie. Przykładowo zmienna dotycząca ogólnej jakości drużyny została utworzona na podstawie średniej liczby punktów tej drużyny w trzech ostatnich sezonach, natomiast zmienna dotycząca formy danej drużyny została utworzona na podstawie średniej liczby punktów zdobytych przez drużynę w pięciu ostatnich meczach.[5]

Na podstawie zmiennej dotyczącej sezonu rozgrywania meczu utworzone zostały również zmienne pomocnicze, takie jak liczba meczów rozegranych uprzednio w danym sezonie przez daną drużynę, liczba meczów domowych rozegranych uprzednio w danym sezonie przez daną drużynę, zmienna informująca o tym, od jak dawna dana drużyna uczestniczy w meczach Bundesligi i tym podobne. Zmienne te służą do dokonania operacji na liczbie punktów zdobywanych przez drużyny.

Drugi utworzony przeze mnie zbiór danych dotyczy kursów bukmacherskich od sezonu 2004/2005 do 2018/2019 dla czterech bukmacherów: Bet365 (B365), Bet&Win (BW), Interwetten (IW) oraz William Hill (WH). Dodatkowo dla każdego sezonu wyliczony został procent rezultatów prawidłowo przewidzianych przez bukmachera[6]

Ostatecznie zbiór danych wykorzystanych przeze mnie w prognozowaniu wyników meczów przyjął postać ukazaną w tabeli 1.1.

1. *Tabela 1.1 Opis zmiennych*

Zmienna	Opis	Wartości zmiennej
Date	Data meczu	Od 1993-08-07 do 2019-03-31
Season_Years	Sezon	Tekst (od 1993/1994 do 2018/2019)
Season	Numer sezonu od 1993/94	Numeryczna (od 1 do 26)
Matchday_Home	Który z kolei mecz gra drużyna domowa w danym sezonie (część meczów była przełożona, więc wartość nie musi być równa kolejce)	Numeryczna (od 1 do 34)
Matchday_Away	Który z kolei mecz gra drużyna wyjazdowa w danym sezonie	Numeryczna (od 1 do 34)
Home_Team	Nazwa drużyny domowej	Zmienna tekstowa

2. *Tabela 2.1 Opis zmiennych, cd*

Zmienna	Opis	Wartości zmiennej
Away_Team	Nazwa drużyny wyjazdowej	Zmienna tekstowa
Result	Wynik meczu (zmienna objaśniana)	Kategoryczna (Zwycięstwo drużyny domowej - H, wyjazdowej - A, remis - D)
Matchday_At_Home_Home	Który z kolei mecz u siebie gra drużyna domowa w danym sezonie	Numeryczna (od 1 do 17)
How_Long_Home	Od jak dawna drużyna domowa gra już w Bundeslidze	Kategoryczna (Expansion – beniaminek[7], 1-season – od jednego sezony, 2-seasons – od dwóch sezonów, Long – 3 lub więcej sezonów)
All_Seasons_Home	Czy drużyna domowa grała w Bundeslidze przez wszystkie sezony od 1993 do 2018	Kategoryczna (Tak lub Nie)
Pts_Home	Liczba punktów w danym meczu drużyny domowej	Numeryczna (0, 1, lub 3)
Home_Team_Quality	Jakość drużyny domowej mierzona jako średnia liczba punktów na mecz z trzech ostatnich sezonów	Numeryczna (od 0 do 3)
Home_Team_Form	Forma drużyny domowej mierzona jako łączna liczba punktów zdobytych przez drużynę w pięciu ostatnich meczach	Numeryczna (od 0 do 15)
Current_Table_Home	Pozycja w tabeli drużyny domowej mierzona średnią liczbą punktów na mecz w obecnym sezonie	Numeryczna (od 0 do 3)

3. *Tabela 3.1 Opis zmiennych, cd*

Zmienna	Opis	Wartości zmiennej
Home_Team_Home_Form	Forma drużyny domowej u siebie mierzona jako łączna liczba punktów w pięciu ostatnich meczach rozgrywanych na własnym stadionie	Numeryczna (od 0 do 15)
Avg_2seasons_Direct_Home	Średnia liczba punktów na mecz zdobyta w czterech poprzednich meczach przez drużynę domową z daną drużyną wyjazdową	Numeryczna (od 0 do 3)
Is_BBR_Home	Czy drużyna domowa jest Bayernem, Borussią lub Lipskiem	Kategoryczna (0 – Nie, 1 – Tak)
Number_10D_Home	Liczba remisów drużyny domowej w dziesięciu ostatnich meczach	Numeryczna (od 0 do 10)
Matchday_Away_Away	Który z kolei mecz na wyjeździe gra drużyna wyjazdowa w danym sezonie	Numeryczna (od 1 do 17)
How_Long_Away	Od jak dawna drużyna wyjazdowa gra już w Bundeslidze	Patrz: How_Long_Home
All_Seasons_Away	Czy drużyna wyjazdowa grała w Bundeslidze przez wszystkie sezony od 1993 do 2018	Kategoryczna (Tak lub Nie)
Pts_Away	Liczba punktów w danym meczu drużyny wyjazdowej	Numeryczna (0, 1, lub 3)
Away_Team_Quality	Jakość drużyny wyjazdowej mierzona jako średnia liczba punktów na mecz z trzech ostatnich sezonów	Numeryczna (od 0 do 3)
Away_Team_Form	Forma drużyny wyjazdowej mierzona jako łączna liczba punktów zdobytych przez drużynę w pięciu ostatnich meczach	Numeryczna (od 0 do 15)

4. Tabela 4.1 Opis zmiennych, cd

Zmienna	Opis	Wartości zmiennej
Current_Table_Away	Pozycja w tabeli drużyny wyjazdowej mierzona średnią liczbą punktów na mecz w obecnym sezonie	Numeryczna (od 0 do 3)
Away_Team_Away_Form	Forma drużyny wyjazdowej na wyjeździe mierzona jako łączna liczba punktów w pięciu ostatnich meczach rozgrywanych na obcym stadionie	Numeryczna (od 0 do 15)
Is_BBR_Away	Czy drużyna wyjazdowa jest Bayernem, Borussią lub Lipskiem	Kategoryczna (0 – Nie, 1 – Tak)
Number_10D_Away	Liczba remisów drużyny wyjazdowej w dziesięciu ostatnich meczach	Numeryczna (od 0 do 10)

Przed wybraniem zmiennych do budowy modelu należało się zastanowić, jakie czynniki mogą wpływać na rezultat pojedynczego meczu. Z mojej wieloletniej obserwacji meczów piłkarskich wynika, że:

1. Istnieją drużyny będące faworytami ze względu na wysoką jakość gry prezentowaną długoterminowo, w poprzednich sezonach. Z tego powodu utworzone zostały zmienne „Home_Team_Quality" oraz „Away_Team_Quality" bazujące na punktach zdobytych przez drużynę w trzech ostatnich sezonach.
2. Bywa tak, że dany zespół nawet jeśli nie jest zaliczany do najlepszych zespołów ligi, seryjnie wygrywa kilka meczów pod rząd. Mówi się wtedy o wysokiej formie danego zespołu i jego szanse na wygraną w kolejnym meczu również rosną. Analogicznie istnieją zespoły zaliczane do najlepszych w lidze, ale będące w kiepskiej formie (na przykład Schalke 04 z obecnego sezonu). Zmienne „Home_Team_Form" oraz „Away_Team_Form" pozwalają na wychwycenie tych zależności na podstawie liczby punktów zdobytych przez drużynę w pięciu ostatnich meczach.
3. Istnieją drużyny, które wybitnie preferują grę na własnym stadionie a wybitnie nie lubią grać na wyjeździe. Na wychwycenie takich zależności pozwolą zmienne

„Home_Team_Home_Form" oraz „Away_Team_Away_Form" zbudowane na podstawie liczby punktów danej drużyny w pięciu ostatnich meczach domowych (wyjazdowych).

4. Im wyżej drużyna w tabeli ligowej w danym momencie, tym wyższe prawdopodobieństwo wygranej w nadchodzącym meczu. Stąd utworzone zostały zmienne „Current_Table_Home" oraz „Current_Table_Away" obrazujące pozycję drużyny w tabeli na podstawie średniej liczby punktów na mecz zdobytych od początku sezonu.
5. Istnieją drużyny, które zaliczają szczególnie dużo remisów lub zaliczają bardzo długie serie kolejnych remisów. Aby wychwycić takie drużyny, utworzone zostały zmienne „Number_10D_Home" oraz „Number_10D_Away" wskazujące liczbę remisów danego zespołu w ostatnich dziesięciu meczach.
6. Niektóre drużyny wybitnie nie lubią grać przeciwko innej, konkretnej drużynie. Na przykład Borussia Dortmund wygrała przeciwko Herthcie Berlin zaledwie dwa z ostatnich sześciu meczów Bundesligi, mimo że generalnie kończy sezon na zdecydowanie wyższym miejscu w tabeli niż Hertha. Stąd utworzona została zmienna „Avg_2seasons_Direct_Home" obrazująca średnią liczbę punktów na mecz zespołu domowego przeciwko konkretnej drużynie wyjazdowej w dwóch ostatnich sezonach Bundesligi.

Powyższe związki opierają się na moich obserwacjach i zostaną empirycznie zbadane w podrozdziale *Korelacje.*

Braki danych

Głównym problemem, który pojawił się podczas tworzenia zmiennych, były braki danych. Metodami, które zostały zastosowane w celu poradzenia sobie z tym problemem były: usunięcie obserwacji zawierających braki danych oraz uzupełnianie średnią.

Pierwsza metoda zastosowana została do trzech pierwszych sezonów w zbiorze danych. Ponieważ zmienne „Home_Team_Quality" oraz „Away_Team_Quality" wymagają obserwacji dotyczących punktów zdobytych przez daną drużynę w meczach z trzech ostatnich sezonów, dla sezonów 1993/94, 1994/95 i 1995/96 takie dane nie były dostępne. Sezony te zostały zatem usunięte ze zbioru danych.

Metoda uzupełniania średnią została wykorzystana w pozostałych sezonach. Niestety, tylko 6 drużyn występujących w zbiorze danych brało udział w Bundeslidze przez wszystkie sezony

od 1996/97 do 2017/18. Dla pozostałych drużyn na pewnym etapie musiała zatem wystąpić konieczność uzupełnienia braków danych. Ogólną ideę uzupełnienia braków danych można pokazać na przykładzie zmiennych „Home_Team_Quality" oraz „Away_Team_Quality". Jak już wspomniałem bazują one na średniej liczbie punktów na mecz danej drużyny zdobytych w trzech ostatnich sezonach w Bundeslidze. Dla drużyn, które w danym punkcie w czasie grały w Bundeslidze od mniej niż trzech lat musiały zatem wystąpić braki danych. Drużyny takie zostały podzielone na trzy grupy:

- Grupę beniaminków, czyli drużyn, które awansowały w danym sezonie do najwyższej ligi – braki danych dla tych drużyn zostały uzupełnione średnią liczbą punktów na mecz drużyn, które były beniaminkami w trzech ostatnich sezonach Bundesligi. Dla przykładu, w sezonie 2018/19 jednym z beniaminków była Fortuna Düsseldorf. Natomiast beniaminkami z trzech ostatnich sezonów były: w sezonie 2015/16 – Ingolstadt i Darmstadt, w sezonie 2016/17 – Lipsk i Freiburg, w sezonie 2017/18 - Stuttgart i Hannover. Obserwacje dotyczące jakości zespołu Fortuny Düsseldorf w sezonie 2018/19 zostały zatem wyliczone jako średnia liczba punktów na mecz wspomnianych drużyn we wspomnianych sezonach.
- Grupę drużyn, które grały w Bundeslidze drugi sezon z rzędu – braki danych dla tych drużyn zostały uzupełnione jako średnia z liczby punktów danej drużyny w poprzednim sezonie oraz liczby punktów beniaminków z dwóch sezonów przed poprzednim sezonem
- Grupę drużyn, które grały w Bundeslidze trzeci sezon z rzędu – braki danych dla tych drużyn zostały uzupełnione jako średnia z liczby punktów danej drużyny z dwóch ostatnich sezonów oraz liczby punktów beniaminków z sezonu poprzedzającego dwa poprzednie sezony

Braki danych dla innych zmiennych zostały zakodowane w podobny sposób. Dla zmiennych „Home_Team_Form", „Home_Team_Home_Form" i kilku innych, wystąpiła dodatkowo konieczność poradzenia sobie z faktem, że nie mają one sensu dla pierwszych pięciu kolejek w każdym sezonie. Przy wyliczaniu formy drużyny w pierwszych meczach danego sezonu (które odbywają się zazwyczaj w sierpniu), bezcelowe jest branie pod uwagę meczów z ostatnich kolejek sezonu poprzedniego (które odbywają się zazwyczaj w maju) ze względu na dystans czasowy je dzielący. Dla pierwszych pięciu kolejek każdego sezonu braki zostały uzupełnione jako średnia liczba punktów na mecz danej drużyny w poprzednim sezonie pomnożona przez 5. Braki danych dotyczące beniaminków zostały uzupełnione w sposób analogiczny do zmiennej „Home_Team_Quality".

Celem budowanego przeze mnie modelu jest przewidzenie wszystkich meczów Bundesligi niezależnie od numeru kolejki oraz faktu, czy dana drużyna jest beniaminkiem, czy też nie. Gdyby celem modelu było przewidzenie tylko wyników meczów, dla których obserwacje nie zawierałyby braków danych (a zatem wyłącznie meczów kolejki dalszej niż piąta i wyłącznie dla sześciu drużyn, które występowały w Bundeslidze przez wszystkie sezony), jakość modelu byłaby znacznie wyższa. Z drugiej jednak strony liczba obserwacji drastycznie by się zmniejszyła.

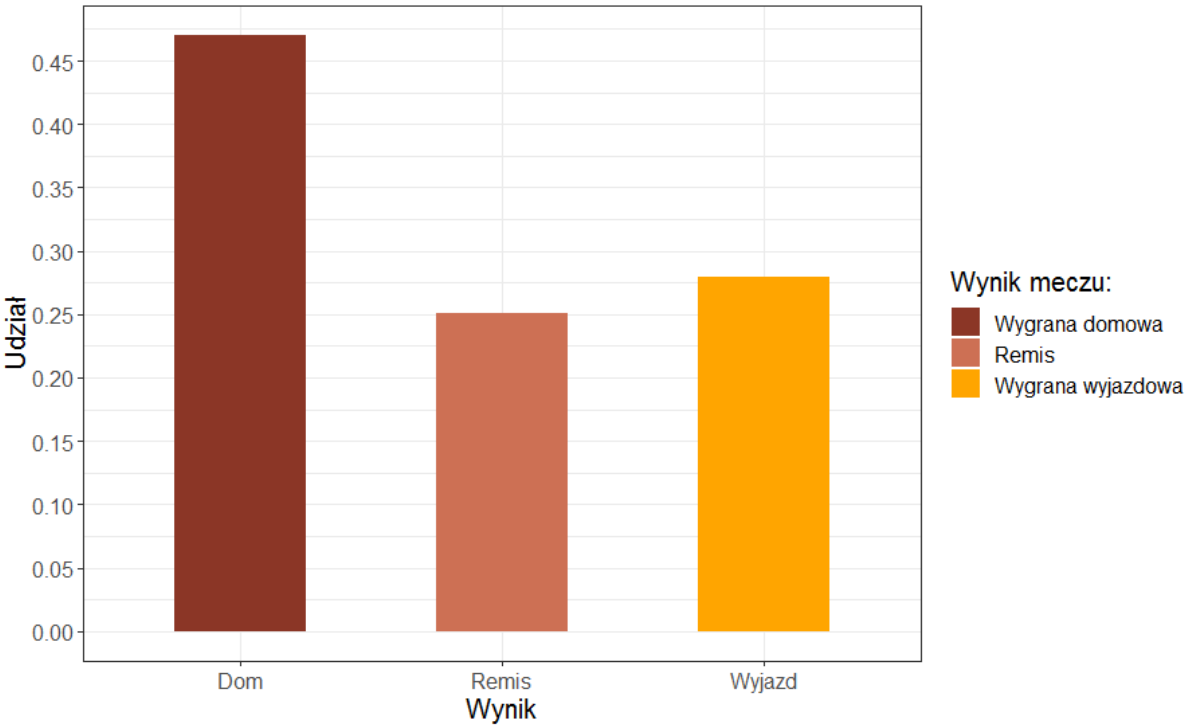
Analiza i wizualizacja danych

W podrozdziale tym spróbuję zilustrować ogólny charakter zbioru danych oraz uzasadnić, dlaczego do budowy modelu wykorzystałem właśnie zmienne wymienione w podrozdziale „Opis zmiennych oraz przekształceń dokonanych na zmiennych". Analizy będą dotyczyły wszystkich w pełni rozegranych sezonów w zbiorze danych, a zatem sezonów od 1996/97 do 2017/18. Sezony od 1993/94 do 1995/1996 zostały usunięte ze względu na braki danych, natomiast sezon 2018/2019 w momencie pisania pracy jeszcze nie dobiegł końca. Łączna liczba zespołów, które grały w Bundeslidze w powyższych latach wyniosła 38. W przypadku, gdy będzie istniała potrzeba porównań między drużynami na wykresach, będą one ze względu na czytelność dotyczyły tylko drużyn, które przez wszystkie sezony od 1996/97 do 2017/2018 grały w Bundeslidze. Jest to sześć drużyn: Bayern Monachium, Borussia Dortmund, Schalke 04, Bayer Leverkusen, Werder Brema oraz Hamburger SV (który spadł do 2. Bundesligi dopiero po sezonie 2017/18).

Zależności i statystyki

Analizę warto zacząć od zbadania rozkładu zmiennej objaśnianej. Zmienna objaśniana ma postać trzech kategorii. Literka H oznacza zwycięstwo w danym meczu drużyny domowej, literka D – remis, literka A - zwycięstwo drużyny wyjazdowej. Intuicja podpowiada, że zwycięstwa drużyny domowej są znacznie częstsze od zwycięstw drużyny wyjazdowej. Drużyny z zasady preferują grę na własnym stadionie, wśród własnych kibiców, którzy często niosą drużynę do zwycięstwa. Wykres 1.1 obrazuje procentowy rozkład zwycięstw drużyny domowej, zwycięstw drużyny wyjazdowej i remisów.

Wykres 1.1 Struktura wyników w Bundeslidze

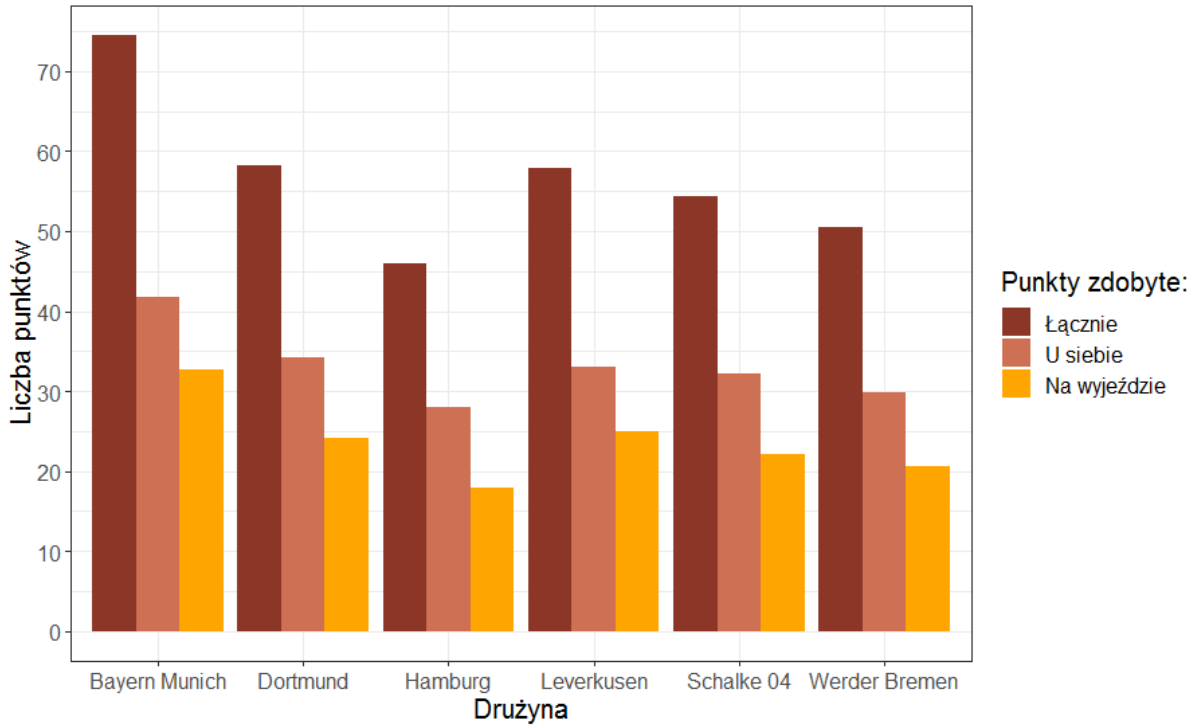


Z wykresu i analiz wynika, że w sezonach 1996 /1997 do 2017/2018 około 47% meczów zostało wygranych przez drużyny grające na własnym stadionie. Zwycięstwa wyjazdowe stanowiły około 28% przypadków. Najrzadsze natomiast były remisy, które pojawiały się w 25% meczów. Istnieje zatem wyraźna dysproporcja między kategorią H a kategoriami A i D, jednak można uznać wszystkie klasy za wystarczająco liczne.

O fakcie, że drużyny wolą grać na własnym stadionie świadczyć może również wykres 1.2. Z wykresu wyraźnie widać, że każdy z szóstki najbardziej stabilnych zespołów w Bundeslidze preferuje grę przed własną publicznością. W ujęciu procentowym, z zespołów ukazanych na wykresie, największą różnicę miejsce rozgrywania meczu robiło Hamburgowi, dla którego punkty zdobyte u siebie stanowiły około 61% łącznej liczby punktów. Ze wszystkich 38 drużyn, które grały w Bundeslidze w latach 1996-2018 największą dysproporcję odnotowano dla zespołu Unterhaching (78,5% punktów to punkty zdobyte u siebie), który jednak grał w najwyższej klasie rozgrywkowej jedynie w sezonach 1999/2000 i 2000/2001. Z zespołów, które można uznać za bywalców Bundesligi najwyższą preferencję gry u siebie odnotowano dla Borussii Monchengladbach – 67%. Po drugiej stronie znajdują się zespoły Bayernu Monachium i Bayeru Leverkusen (tylko 56-57% punktów to punkty zdobyte u siebie). Najniższy procent, 19%, odnotowano dla drużyny Greuter Furth, która w swoim jedynym sezonie 2012/2013 zdobyła u siebie zaledwie 4 punkty. Jest to również jedyna drużyna

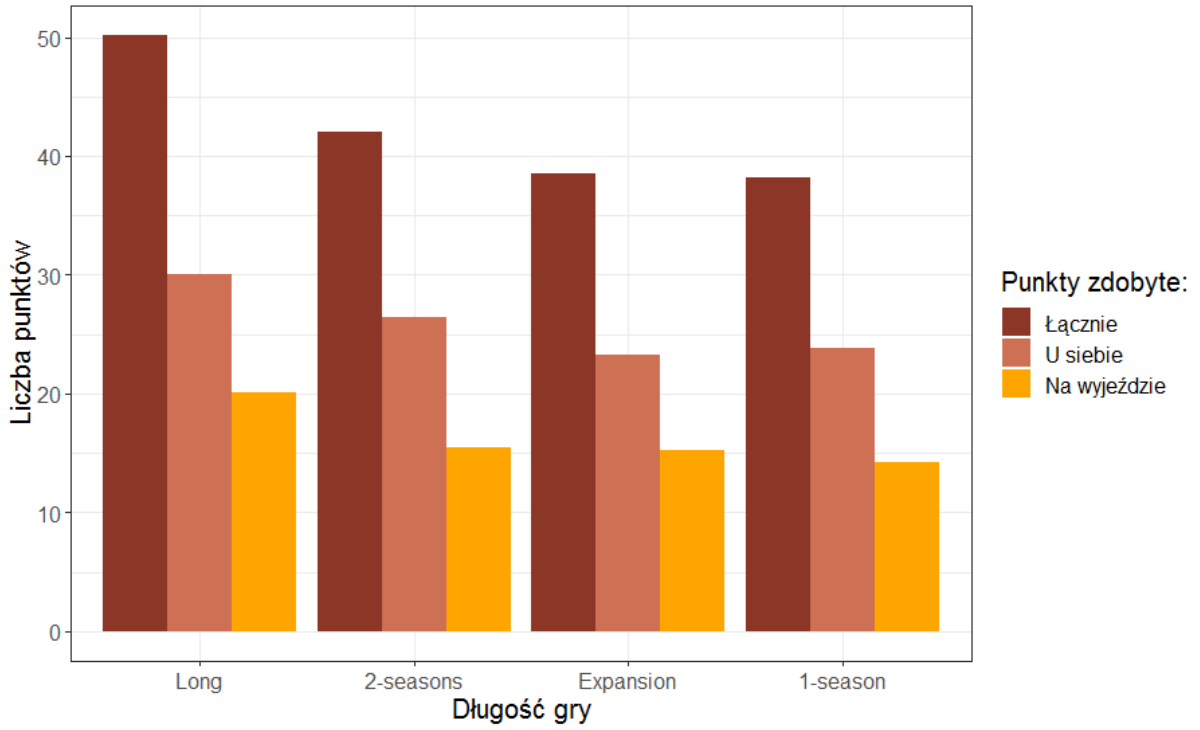
w zbiorze, która średnio preferowała grę na wyjeździe nad grą u siebie. Z wykresu i analiz można również wysnuć inne ciekawe wnioski. Na pierwszym miejscu pod względem średniej liczby punktów znalazł się oczywiście Bayern Monachium – 74,5 punktu na sezon. Drugi był Rb Lipsk, który w Bundeslidze gra od dwóch sezonów i średnio zdobywał 60 punktów na sezon. Trzecia natomiast była Borussia Dortmund – średnio 58 punktów o ledwie ułamki punktów wyprzedzając czwarty Bayer Leverkusen.

Wykres 1.2 Liczba punktów zdobywanych średnio przez drużyny, które w latach 1996-2018 nieprzerwanie grały w Bundeslidze



Warto również zbadać, jak przedstawia się liczba punktów zdobywanych przez zespoły ze względu na długość gry w Bundeslidze. Wykres 1.3 obrazuje tę zależność. Zgodnie z intuicją, zespoły, które grają w Bundeslidze od trzech lat lub dłużej (Long) osiągają znacznie wyższe dorobki punktowe. Nie ma jednak znaczącej różnicy pomiędzy drużynami o stażu jednego sezonu a beniaminkami. Wygląda na to, że najtrudniej jest utrzymać się w Bundeslidze przez pierwsze dwa lata, trzeci sezon (zmienna 2-seasons) zazwyczaj wygląda pod względem liczby punktów znacznie lepiej. Po raz kolejny ujrzeć można różnicę pomiędzy preferencją gry u siebie i na wyjeździe. Największą różnicę ten fakt robi drużynom o stażu dwuletnim, najmniejszą beniaminkom.

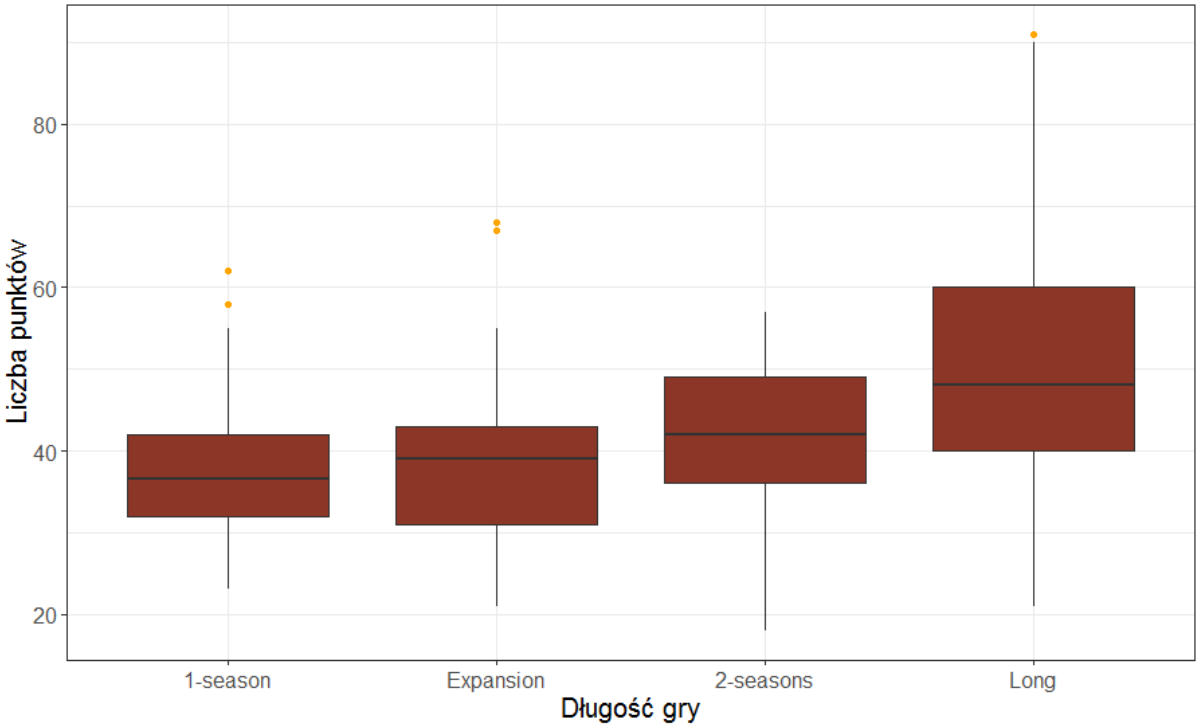
Wykres 1.3 Liczba punktów ze względu na długość gry w Bundeslidze



Obserwacje nietypowe

Piłka nożna, często określana jako sport nieprzewidywalny, potrafi zaskoczyć niespodziewanymi rozstrzygnięciami. Ostatnio najgłośniej było o mistrzowskim sezonie Leicester City w Anglii. Jednak w sezonie 1997/1998 drużyna Kaiserslautern dokonała rzeczy jeszcze bardziej niespodziewanej, a mianowicie zdobyła mistrzostwo Bundesligi jako beniaminek. Ten fakt jest zupełnie niezgodny z analizami dokonanymi na poprzednim wykresie. Wykres „Obserwacje nietypowe ze względu na długość gry w Bundeslidze" pozwala na wykrycie obserwacji nietypowych. Na wykresie zostały one zaznaczone żółtymi kropkami. Dwie kropki dla drużyn grających w Bundeslidze od jednego sezonu i jedna dla drużyn grających od trzech lub więcej sezonów zostały pominięte ze względu na nieduże odchylenie od reszty obserwacji. Warto jednak bliżej przyjrzeć się dwóm obserwacjom nietypowym dla beniaminków. Są to wspomniany Kaiserslautern z sezonu 97/98, kiedy zespół zdobył 68 punktów, oraz Rb Lipsk z sezonu 2016/17 (67 punktów), kiedy to po awansie do Bundesligi bogaty sponsor, RedBull, dokonał transferów za olbrzymie pieniądze.

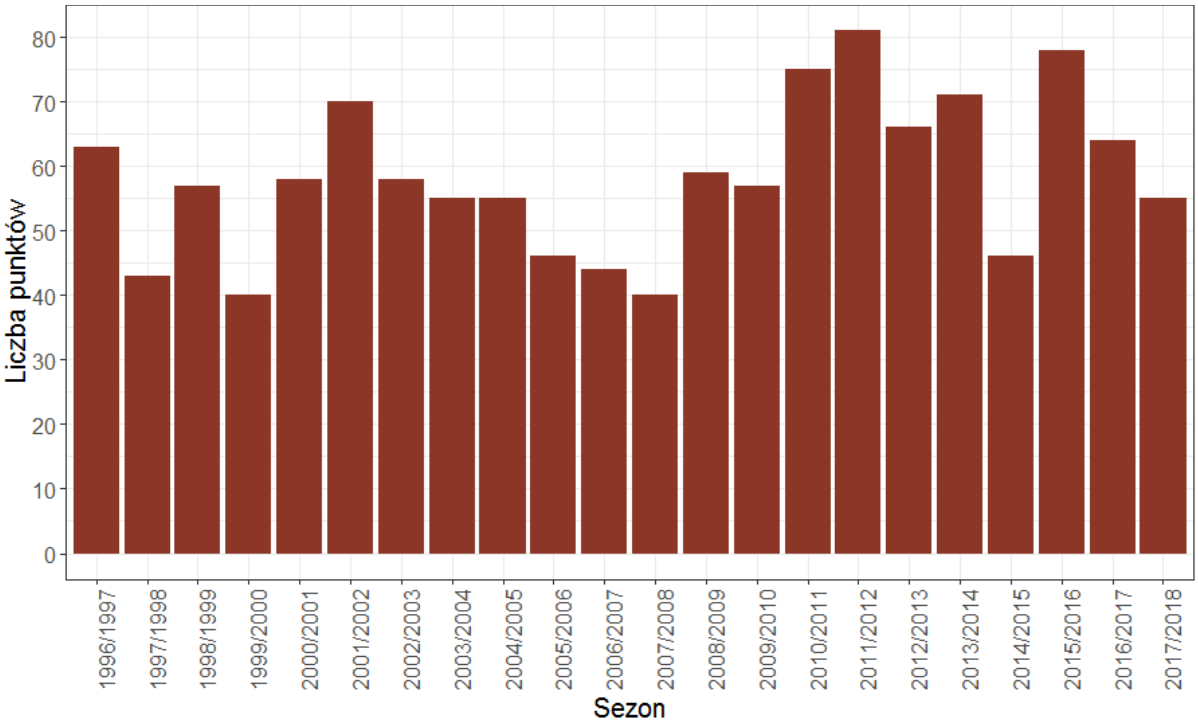
Wykres 1.4 Obserwacje nietypowe ze względu na długość gry w Bundeslidze



Ze względu na bardzo małe prawdopodobieństwo tak wysokiej liczby punktów dla beniaminka, zdecydowałem się nie uwzględniać meczów zespołu Kaiserslautern z sezonu 97/98 w modelu. W tym celu usunąłem obserwacje dotyczące zespołu w mistrzowskim sezonie. Z wyczynem Rb Lipska zdecydowałem się nie robić podobnej rzeczy i obserwacje dla tego zespołu nie zostały usunięte (wyjaśnienie tej decyzji przedstawiam w rozdziale trzecim).

Jako fan Borussii Dortmund nie byłbym sobą, gdybym nie skupił się przynajmniej na chwilę w swojej analizie właśnie na zespole z Zagłębia Ruhry. Pamiętam słaby sezon 2014/2015 w wykonaniu mojej ulubionej drużyny. Pamiętam jak po kilkunastu kolejkach Borussia znajdowała się w strefie spadkowej i pierwszy raz od dawna wszyscy zaczęli mówić o klubie w kontekście walki o utrzymanie w Bundeslidze a nie europejskich pucharów. Główną przyczyną słabej gry była plaga kontuzji wśród obrońców drużyny, co spowodowało w trakcie sezonu konieczność podpisania kontraktu z Manuelem Friedrichem, zawodnikiem, który od roku pozostawał bez klubu. Jako że nie specjalnie interesowałem się wtedy statystykami wydawało mi się, że musi to być jeden z najgorszych sezonów w historii klubu. Na wykresie 1.5 zweryfikowane zostały te przypuszczenia.

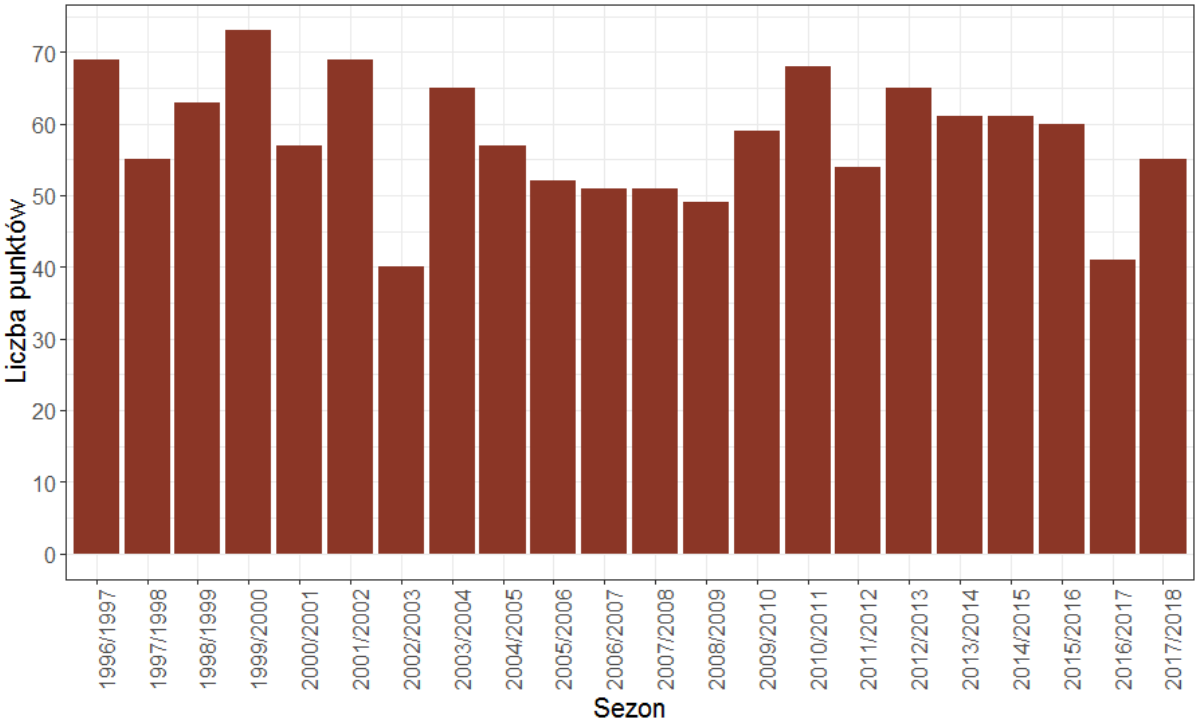
Wykres 1.5 Liczba punktów Borussii Dortmund na przestrzeni lat



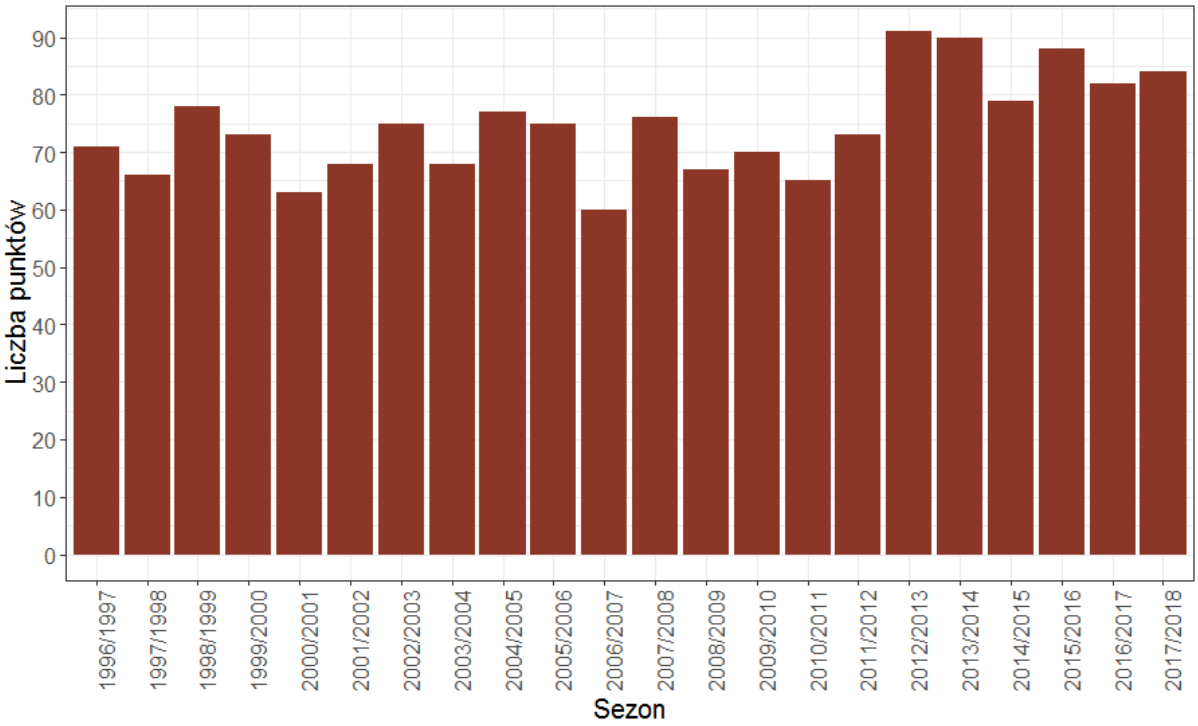
Kiedy spojrzy się 10 lat wstecz, to wynik Borussii z sezonu 2014/2015 nie wydaje się specjalnie zaskakujący. Zdarzały się sezony, w których zespół notował gorszy dorobek punktowy. Jednak sezon 2014/15 był dla klubu znacznie gorszy od sezonów bezpośrednio go poprzedzających i sezonu kolejnego. Można stwierdzić z przekonaniem, że przed sezonem nikt nie spodziewał się podobnego obrotu spraw i był to tylko „wypadek przy pracy".

Jak wcześniej zostało ustalone, obok Borussii dwa najlepsze kluby w historii Bundesligi od 1996 roku, to Bayern Monachium i Bayer Leverkusen (pomijając RB Lipsk, który gra w Bundeslidze dopiero trzeci sezon). Te trzy drużyny w większości rozgrywanych meczów Bundesligi są typowane jako faworyci do zwycięstwa i każdy inny rezultat można określić jako niespodziewany. Warto sprawdzić, czy również dla drużyn Bayernu i Bayeru istniały sezony, w których te drużyny zagrały znacznie poniżej oczekiwań. Wykresy 1.6 i 1.7 pokazują liczbę punktów Bayeru i Bayernu na przestrzeni lat.

Wykres 1.6 Liczba punktów Bayeru Leverkusen na przestrzeni lat



Wykres 1.7 Liczba punktów Bayernu Monachium na przestrzeni lat



W przypadku Bayeru Leverkusen wyraźnie można dostrzec dwa słabsze sezony: 2002/2003 i 2016/2017. Wykres liczby punktów Bayernu Monachium pokazuje natomiast,

dlaczego to właśnie Bayern może być uważany za najlepszy zespół w historii Bundesligi. Klub z Bawarii niemal nie ma słabszych sezonów, zawsze gra na stabilnym i bardzo wysokim poziomie. Ciężko jest dostrzec tutaj jakiekolwiek obserwacje nietypowe.

Korelacje

Kolejnym etapem analizy będzie zbadanie korelacji między rezultatem meczu a poszczególnymi zmiennymi w celu określenia, które zmienne faktycznie mogą mieć wpływ na wartość zmiennej objaśnianej. Tabela 1.2 przedstawia średnie wartości poszczególnych zmiennych dla poszczególnych rezultatów meczu:

Tabela 1.2 Średnia wartość zmiennych dla poszczególnych rezultatów meczu

	Result*		
Zmienne objaśniające	**A**	**D**	**H**
Home_Team_Quality	1.321	1.368	1.461
Away_Team_Quality	1.493	1.398	1.343
Number_10D_H	2.600	2.617	2.520
Number_10D_A	3.446	3.621	3.652
Home_Team_Form	6.064	6.609	7.221
Away_Team_Form	7.567	7.094	6.697
Current_Table_H	1.256	1.340	1.469
Current_Table_A	1.484	1.388	1.305
Avg_2seasons_D_H	1.235	1.340	1.472
Home_Team_H_F	7.735	8.134	8.829
Away_Team_A_F	6.001	5.529	5.134

** A – Zwycięstwo drużyny wyjazdowej, D – Remis, H – zwycięstwo drużyny domowej*

Większość wyników uzyskanych w tabeli jest zgodna z intuicją. Im wyższa jakość drużyny domowej (wyjazdowej), tym wyższe prawdopodobieństwo jej zwycięstwa. Analogicznie, im lepsza forma drużyny domowej (wyjazdowej) czy to ogólna, czy to w meczach domowych (wyjazdowych), tym wyższa szansa na zwycięstwo. Również zmienna dotycząca rezultatów czterech poprzednich meczów pomiędzy tymi samymi drużynami

(„Avg_2seasons_Direct_Home”) silnie i zgodnie z intuicją różnicuje wyniki spotkań. Dwie zmienne dotyczące liczby remisów w dziesięciu ostatnich meczach nie okazały się natomiast dobrymi predyktorami rezultatu meczu. Tworząc te zmienne miałem nadzieje, że duża liczba remisów danej drużyny w poprzednich meczach będzie powodowała wysokie prawdopodobieństwo wystąpienia remisu w meczu obecnym. Niestety, zgodnie z analizami taka zależność nie występuje (powodem braku zależności może być również konieczność uzupełnienia dużej ilości braków danych, co może zaburzać ogólny wynik).

Kolejnym krokiem będzie zbadanie korelacji pomiędzy wszystkimi parami zmiennych objaśniających. Zgodnie z portalem pogotowiestatystyczne.pl[8], za silnie skorelowane można uznać zmienne, dla których współczynnik korelacji przekracza 0,5. Zmiennymi najsilniej skorelowanymi z pozostałymi zmiennymi okazały się zmienne dotyczące pozycji w tabeli mierzonej na podstawie średniej liczby punktów („Current_Table_Home” i „Current_Table_Away”). Zmienne te były silnie skorelowane ze wszystkimi zmiennymi dotyczącymi formy i jakości drużyny. Silne korelacje wystąpiły również pomiędzy zmiennymi dotyczącymi ogólnej formy drużyn („Home_Team_Form”,”Away_Team_Form”) a formą drużyn w meczach domowych i wyjazdowych („Home_Team_Home_Form”, „Away_Team_Away_Form”).

Po analizach dokonanych powyżej, można przystąpić do budowy modelu przewidującego wyniki meczów piłkarskich Bundesligi oraz szacującego prawdopodobieństwa każdego z trzech zdarzeń: wygrana drużyny domowej, wygrana drużyny wyjazdowej lub remis.

Rozdział 2 – Opis metody badawczej i budowa modelu

Poniższy model może zostać zastosowany tylko i wyłącznie do przewidywania wyników najbliższej kolejki Bundesligi, gdyż w celu określenia wartości zmiennych dla danej obserwacji potrzebna jest znajomość wyników kolejki poprzedniej. Model nie może zatem zostać wykorzystany przykładowo do jednoczesnej predykcji wyników meczów na cały nadchodzący sezon. Po oszacowaniu parametrów modelu i ocenie jakości jego predykcji dokonane zostanie porównanie z wynikami modeli czterech wielkich firm bukmacherskich

Zbiór danych i selekcja zmiennych do budowy modelu

W celu umożliwienia porównań z firmami bukmacherskimi, model zostanie wytrenowany na historycznych danych dotyczących meczów Bundesligi z sezonów od 1996/97 do 2015/2016. Zbiorem walidacyjnym, umożliwiającym sprawdzenie jakości modelu i dokonanie porównań, będzie zbiór dotyczący meczów Bundesligi od sezonu 2016/17 do 27. kolejki sezonu 2018/2019.

Do treningu modelu wykorzystanych zostanie 7 zmiennych: "Avg_2seasons_Direct_Home", „Home_Team_Quality",„Away_Team_Quality",„Home_Team_Form",„Away_Team_Form", „Home_Team_Home_Form", "Away_Team_Away_Form". Zmienne "Current_Table_Home" i "Current_Table_Away" zostały z modelu usunięte ze względu na silną korelację z pozostałymi zmiennymi wykrytą w podrozdziale „*Korelacje"*. Zmienne dotyczące liczby remisów „Number_10D_Home" i „Number_10D_Away" zostały usunięte ze względu na brak skorelowania ze zmienną objaśnianą, czyli zmienną „Result".

W podrozdziale „Obserwacje nietypowe", zostało wykrytych kilka grup obserwacji, które można uznać za niespodziewane, i które mogłyby zaburzać trening modelu. W związku z tym ze zbioru zostały usunięte następujące obserwacje:

1. Mecze Kaiserslautern z sezonu 1997/98, kiedy drużyna ta została mistrzem kraju jako beniaminek.
2. Mecze Borussii Dortmund z sezonu 2014/2015, kiedy drużyna zdobyła zaledwie 45 punktów i przez sporą część sezonu broniła się przed spadkiem do niższej ligi.
3. Mecze dotyczące Bayeru Leverkusen z sezonu 2002/2003, kiedy drużyna zdobyła w sezonie zaledwie 40 punktów.

Mecze Lipska oraz Bayeru Leverkusen z sezonu 2016/2017 wykryte wcześniej jako obserwacje nietypowe nie zostaną z modelu usunięte ze względu na znajdowanie się w zbiorze testowym. Na pewno wyniki tych dwóch drużyn w sezonie 2016/17 będą bardzo trudne do przewidzenia przez model i można się spodziewać, że na danych z tego sezonu model nie będzie miał

szczególnie wysokiej jakości. Ostateczny zbiór zawiera 7 zmiennych objaśniających rezultat meczu i 6873 obserwacje.

Trening modelu

Kolejnym etapem jest trening modelu na zbiorze dotyczącym meczów z sezonów 1996/1997 do 2015/2016 (łącznie 6018 obserwacji). W tym celu zastosowane zostały dwa algorytmy – las losowy oraz sieci neuronowe. Powodami zastosowania tych dwóch algorytmów były:

- fakt, że głównym celem modelu jest dokonanie właściwych predykcji a nie interpretowalność i oszacowanie parametrów przy poszczególnych zmiennych. Algorytmy lasu losowego i sieci neuronowych cechują się bardzo dużą dokładnością zwracanych wyników. Niestety, dokładność ta okupiona jest długim czasem trenowania modelu.
- chęć dokonania porównań oraz sprawdzenia, czy na którymś etapie treningu modelu obiema metodami nie zostały popełnione błędy (co mogłoby zostać wykryte, w przypadku gdyby jakość obydwu modeli na zbiorze testowym znacznie się różniła).

Schemat postępowania dla obydwu algorytmów będzie następujący:

1. Na początku zostanie zbudowany podstawowy model z siedmioma zmiennymi i losowo dobranymi parametrami. Za pomocą miary „Accuracy" sprawdzona zostanie jakość tak zbudowanego modelu na zbiorze testowym.
2. Kolejnym krokiem będzie znalezienie optymalnego zestawu parametrów modelu. W tym celu zostanie użyta walidacja krzyżowa. Zbiór treningowy zostanie podzielony na 10 części. Każdorazowo zbiór będzie trenowany na 9 z tych części, a jakość, mierzona odsetkiem poprawnie zaklasyfikowanych przypadków, zostanie zbadana na dziesiątej, walidacyjnej części. Proces ten zostanie powtórzony 10 razy, za każdym razem jakość będzie oceniana na innej części zbioru. Cały proces zostanie powtórzony dla każdego zestawu parametrów. Zestaw parametrów, dla którego średnia jakość na części walidacyjnej będzie najwyższa zostanie wybrany w celu wytrenowania ostatecznego modelu.
3. Przedostatnim etapem będzie wytrenowanie modelu z parametrami uzyskanymi z kroku poprzedniego.
4. Na koniec oceniona zostanie za pomocą miary „Accuracy" jakość tak zbudowanego modelu na zbiorze testowym. W celu uśrednienia wyników proces treningu zostanie

powtórzony 10 razy. Średnia jakość uzyskana na zbiorze testowym zostanie wykorzystana do dokonania porównań pomiędzy modelami.

W kolejnych punktach opiszę dokładniej podstawowe cechy i parametry obydwu algorytmów jak również sposób ich implementacji.

Algorytm lasu losowego

Las, jak sama nazwa wskazuje, składa się z drzew. Podobnie jest w przypadku lasu lasowego, który powstaje z połącznia wielu klasyfikacji metodą drzewa decyzyjnego. Schemat postępowania algorytmu lasu losowego jest następujący[9]:

1. Losowanie z powtórzeniami k obserwacji ze zbioru treningowego, czyli tak zwanej próby bootstrapowej. Pozostałe przypadki są przypisywane do zbioru testowego.
2. Próba bootstrapowa jest używana do uczenia pojedynczego drzewa. Dla każdego drzewa wybierane jest losowo x zmiennych ze zbioru zmiennych wejściowych. Tylko one zostaną użyte do treningu, to znaczy drzewo decyzyjne będzie analizowało jedynie zależności pomiędzy tymi wybranymi zmiennymi a zmienną wyjściową.
3. Mierzona jest jakość klasyfikacji na utworzonym w pierwszym punkcie zbiorze testowym. Analizowana jest dokładność predykcji oraz wpływ poszczególnych zmiennych wejściowych na jakość predykcji.
4. Powyższe punkty są powtarzane n razy ze względu na wybraną liczbę drzew decyzyjnych w lesie.
5. Wyniki zwracane przez model są rezultatem uśrednienia wyników zwracanych przez wszystkie drzewa decyzyjne. Cecha ta sprawia, że algorytm lasu losowego nie jest aż tak bardzo wrażliwy na przeuczenie, czyli nadmierne zapamiętanie zależności na danych treningowych bez zdolności uogólnienia na przypadki wcześniej niewidziane. Wartość zmiennej wyjściowej, którą zwróciła największa liczba drzew decyzyjnych, staje się predykcją modelu.

Do treningu metodą lasu losowego wybrałem pakiet R oferujący liczne biblioteki przydatne w modelowaniu predykcyjnym. Podczas treningu metodą lasu losowego należy ustalić pewne parametry. Parametrami, które zostały użyte są:

- num.trees – liczba drzew decyzyjnych użytych do budowy lasu losowego. Zbyt duża liczba może doprowadzić do zbyt długiego czasu trenowania modelu.
- min.node.size - Minimalna liczba obserwacji w węźle. Pojedyncza gałąź drzewa nie będzie dzielona jeśli liczba przypadków po podziale byłaby mniejsza niż ten parametr.

Zbyt niska wartość parametru może powodować przeuczenie modelu, zbyt wysoka jego niedouczenie.

- splitrule - metoda podziału każdego drzewa[10]
- mtry – liczba zmiennych wylosowanych do treningu pojedynczego drzewa.

Na pierwszym etapie został wytrenowany model z bazowymi parametrami:

- num.trees = 500
- min.node.size = 1
- splitrule = „gini”

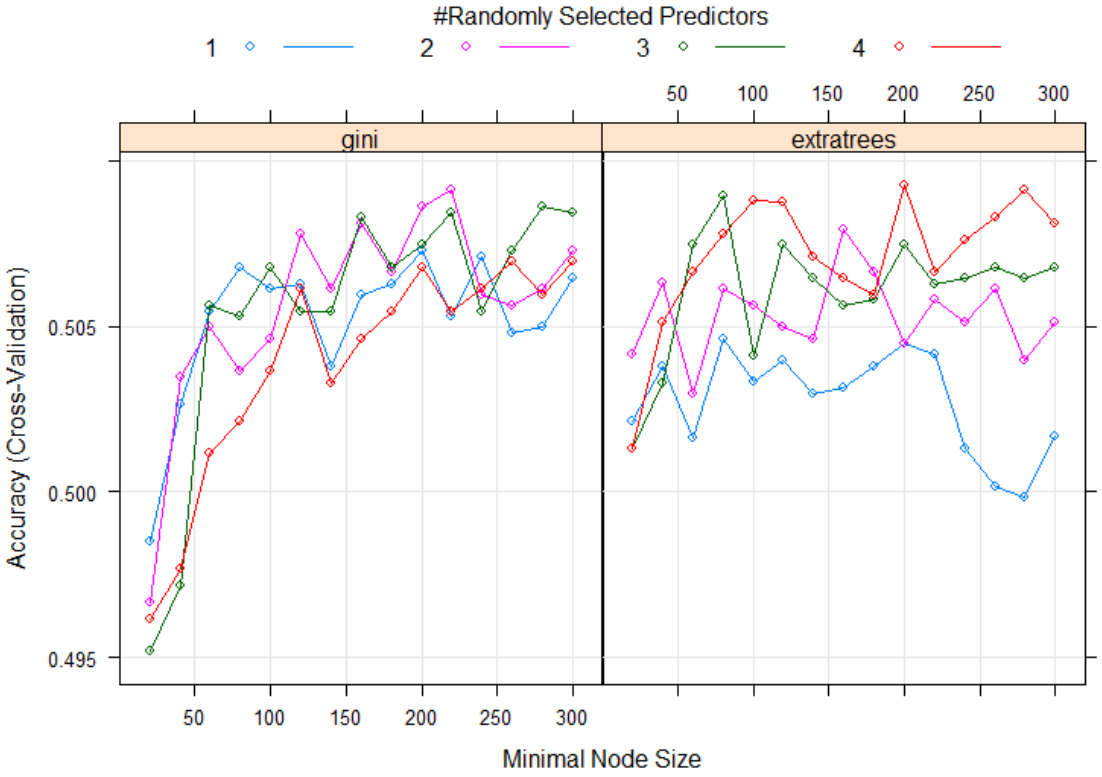
Przy tak dobranych parametrach, na zbiorze treningowym trafność modelu wyniosła 99,7%. Na zbiorze testowym uzyskane zostało 48,9%. Ze względu na olbrzymią różnice w jakości pomiędzy zbiorem treningowym a testowym można uznać, że model przy tak dobranych parametrach został nadmiernie dopasowany do danych w próbie uczącej.

Parametry dobrane na pierwszym etapie okazały się nienajlepszym wyborem. W drugim etapie zostanie dokonany tak zwany „Parameters Tuning”, to znaczy poszukiwanie optymalnych parametrów lasu losowego. Pakiet „caret” umożliwia optymalizację tylko niektórych parametrów. Nie ma możliwości optymalizacji parametru „num.trees” dotyczącego liczby drzew decyzyjnych wykorzystanych do budowy lasu, zatem parametr ten arbitralnie ustawiony został przeze mnie na wartość 150. Optymalizacja parametrów może być dokonana tylko po podaniu określonych wartości parametrów do przeszukania. Ze względu na ograniczenia czasowe i technologiczne, liczba wartości parametrów nie może być zbyt duża. Parametrami, których optymalizacja została dokonana są:

- min.node.size, przeszukiwane wartości: od 20 do 300 co 20 jednostek
- splitrule, przeszukiwane wartości: „gini” oraz „extratrees”
- mtry, przeszukiwane wartości: 1,2,3,4

Jakość predykcji ze względu na zastosowane parametry przedstawiona jest na wykresie 2.1.

Wykres 2.1 Jakość predykcji w zależności od parametrów lasu losowego



Na podstawie 10 – krotnej walidacji krzyżowej dla algorytmu lasu losowego najlepszymi parametrami ze względu uzyskanie najwyższej średniej jakości, 50,76 % na zbiorze walidacyjnym okazały się:

- Minimalna liczba obserwacji w węźle – min.node.size = 200
- Metoda podziału drzewa – splitrule = „extratrees”
- liczba zmiennych użyta do trenowania pojedynczego drzewa – mtry = 4

Ponadto na wykresie, zwłaszcza dotyczącym metody podziału „gini”, widać, na czym polega ryzyko przeuczenia modelu przy zbyt małej wartości parametru min.node.size.

Za pomocą najlepszych parametrów 10 – krotnie został wytrenowany ostateczny model lasu losowego z użyciem 150 drzew decyzyjnych. Każdorazowo jakość modelu została sprawdzona na zbiorze testowym. Średnia trafność uzyskanych prognoz wyniosła 49,95%. Macierz pomyłek na zbiorze testowym przedstawiona została w tabeli 2.1. Ciekawym wnioskiem, jaki można wyciągnąć z tabelki, jest fakt, że model w ogóle nie typuje remisów. Widocznie wynik remisowy jest na tyle przypadkowy, że ciężko znaleźć jakieś wzorce w danych, które wskazywałyby na możliwość wystąpienia takiego rezultatu. Model dla 82 % meczów wytypował zwycięstwo drużyny domowej, natomiast tylko dla 18 % zwycięstwo wyjazdowe. 50,5 % wytypowanych przez model zwycięstw drużyny domowej faktycznie zakończyło się

takim wynikiem, natomiast tylko 48,3 % wytypowanych zwycięstw wyjazdowych zakończyło się rzeczywiście zwycięstwem gościa.

Tabela 2.1 Macierz pomyłek na zbiorze testowym, lasy losowe

	Rzeczywiste rezultaty		
Predykcje	A	D	H
A	73	34	44
D	0	0	0
H	168	181	355

Algorytm sieci neuronowych

Drugim obok lasu drzew decyzyjnych najpotężniejszym algorytmem uczenia maszynowego są sieci neuronowe. Bazują one na budowie ludzkiego mózgu. Ludzki mózg składa się z neuronów wzajemnie ze sobą powiązanych i komunikujących się. Pojedynczy neuron aktywuje się, kiedy sygnał docierający do niego z poprzednich neuronów przekroczy pewien próg (próg aktywacji). Połączenia między neuronami reprezentowane są przez wagi. Wagi są najważniejszym parametrem sieci neuronowej i są one dostosowywane w procesie uczenia sieci. Najprostsza sieć neuronowa składa się z jednej warstwy wejściowej i jednej warstwy wyjściowej. W celu umożliwienia znajdowania przez sieć nieliniowych zależności, do sieci dodaje się również warstwy ukryte, które znajdują się pomiędzy warstwą wejściową a wyjściową. W uproszczeniu proces uczenia sieci neuronowej przebiega następująco:

1. Ustalenie budowy sieci (liczba warstw ukrytych, liczba neuronów w warstwach ukrytych, połączenia między neuronami – domyślnie każdy neuron połączony jest z każdym neuronem z sąsiednich warstw)
2. Losowe inicjowanie wag, czyli połączeń pomiędzy neuronami
3. Obliczanie wartości zmiennej wyjściowej dla danego przypadku treningowego oraz błędu neuronu wyjściowego na podstawie różnicy pomiędzy uzyskaną wartością zmiennej wyjściowej a jej rzeczywistą wartością.
4. Wysyłanie sygnału o uzyskanej różnicy z powrotem do poprzednich warstw.

5. Pobranie kolejnego przypadku treningowego
6. Po pobraniu pewnej, określonej liczby przypadków, aktualizacja wag sieci na podstawie uzyskanych wcześniej błędów neuronów wyjściowych.
7. Po przeanalizowaniu całego zbioru treningowego zmiana kolejności przypadków i rozpoczęcie kolejnej iteracji.
8. Rozpoczęcie całego procesu od nowa.

Do treningu modelu algorytmem sieci neuronowych użyłem Pythonowej biblioteki „Keras". W procesie treningu, użyte zostały przeze mnie następujące parametry sieci neuronowej:

- units – liczba neuronów w warstwie ukrytej. Liczba ta nie powinna być zbyt wysoka ze względu na ryzyko przeuczenia modelu ani zbyt niska ze względu na ryzyko niedouczenia. Przyjmuje się, że początkowa liczba neuronów ukrytych powinna być równa średniej arytmetycznej liczby neuronów w warstwie wejściowej i liczby neuronów wyjściowych.
- kernel_initializer – sposób ustalania początkowych wartości wag pomiędzy neuronami
- activation – funkcja aktywacji, czyli funkcja przekształcająca wartość wejściową danego neuronu (iloczyn wag i wartości neuronów z warstwy poprzedniej) na wartość wyjściową neuronu. Neuron jest aktywowany jeśli funkcja aktywacji przekroczy pewną wartość (tzw. próg aktywacji neuronu).
- optimizer – metoda minimalizacji błędu, czyli różnicy między rzeczywistymi a prognozowanymi wartościami zmiennych wyjściowych
- loss – funkcja straty, czyli funkcja, której wartość jest minimalizowana metodą optymalizacji błędu wskazaną przez parametr optimizer
- metrics – metoda pomiaru jakości modelu
- batch_size – liczba obserwacji pobranych przez sieć neuronową, po której optymalizowane są wagi połączeń
- epochs – liczba epok, czyli przejść całego zbioru treningowego przez sieć neuronową

Spośród powyższych parametrów, następujące przez całą moją analizę pozostają niezmienione:

Dla całego modelu:

- loss="sparse_categorical_crossentropy" – funkcja straty używana w przypadku modelów klasyfikacyjnych z trzema lub więcej kategoriami zmiennej objaśnianej

- metrics=”accuracy” – metodą pomiaru jakości modelu będzie miara trafności predykcji, czyli udział przypadków prawidłowo zaklasyfikowanych
- kernel_initializer = „uniform” – wagi początkowe ustalane są na losowe wartości bliskie 0

Dla warstwy ukrytej:

- units = 5. Ponieważ liczba neuronów w warstwie wejściowej jest równa 7 (liczba zmiennych objaśniających modelu) a liczba neuronów w warstwie wyjściowej jest równa 3 (liczba kategorii zmiennej wyjściowej), więc średnia arytmetyczna neuronów w warstwie wejściowej i wyjściowej równa jest 5, stąd units=5
- activation= „relu” – rectified linear unit function, funkcja aktywacji, która zwraca 0 jeśli wartość wejściowa jest mniejsza lub równa 0 (a więc neuron nie zostaje aktywowany) oraz wartość wejściową, jeśli jest ona większa od 0. Jest to najpopularniejsza funkcja aktywacji.

Dla warstwy wyjściowej:

- units = 3, ponieważ liczba kategorii zmiennej wyjściowej (wygrana domowa, remis, wygrana wyjazdowa) jest równa 3
- activation= „softmax”. Funkcja aktywacji która zwraca prawdopodobieństwo, że zmienna należy do danej klasy w przypadku, gdy klas jest co najmniej 3

Na pierwszym etapie został wytrenowany model z parametrami stałymi podanymi powyżej oraz następującymi parametrami zmiennymi:

- batch_size = 100
- epochs = 100
- optimizer = „adam” – rodzaj metody najszybszego spadku gradientowego

Na tak zbudowanym modelu, dokładność predykcji wyniosła 49,1 % na zbiorze testowym natomiast 51,15 % na zbiorze treningowym. Model można zatem uznać za lekko przetrenowany, co może być spowodowane zbyt dużą liczbą epok, a więc zbyt wysoką wartością parametru „epochs”.

Drugi etap budowy modelu objął znalezienie optymalnych parametrów sieci neuronowej w celu maksymalizacji jakości predykcji. Za pomocą metody „grid search” sprawdzone zostały następujące parametry[11] i ich wartości:

- batch_size, wartości: 20, 40 oraz 60
- epochs, wartości: 30, 50, 70
- optimizer: „adam”, „rmsprop”

Na zbiorach walidacyjnych metodą 10 – krotnej walidacji krzyżowej, najwyższą średnią jakość modelu wynoszącą 50,81 % uzyskano dla parametrów:

- batch_size = 20
- epochs = 30
- optimizer = „adam”

Za pomocą optymalnych parametrów uzyskanych powyżej, model został wytrenowany 10 razy, za każdym razem jego jakość sprawdzona została na zbiorze testowym. Średnia jakość na zbiorze testowym wyniosła 49,91 %. Model typuje zwycięstwo drużyny domowej w 78,6 % przypadków, zwycięstwo drużyny wyjazdowej w 21,4 % przypadków, nie typuje remisów. 50,9% wytypowanych zwycięstw drużyny domowej faktycznie zakończyło się takim rezultatem, w przypadku zwycięstw drużyny wyjazdowej jest to tylko 45,9 %. Macierz pomyłek na zbiorze testowym ukazana jest w tabeli 2.2.

Tabela 2.2 Macierz pomyłek na zbiorze testowym, sieci neuronowe

	Rzeczywiste rezultaty		
Predykcje	A	D	H
A	84	42	57
D	0	0	0
H	157	173	342

Porównanie wyników algorytmu lasu losowego i sieci neuronowej

Główną miarą, którą zdecydowałem się użyć do porównań obu modeli, jest jakość uzyskana na zbiorze testowym, czyli na danych z lat 2016/17 do marca 2018/2019. Dla tego okresu minimalnie lepszy okazał się model wytrenowany metodą lasu losowego, jednak różnica w trafnościach 49,95% do 49,91 % może mieć charakter przypadkowy. Oba modele charakteryzowały się również bardzo zbliżoną wariancją oszacowań. Co ciekawe algorytm lasów losowych nieco częściej typował zwycięstwa drużyny domowej (82 %, w przypadku sieci neuronowej 78,6 %), ale miał niższą skuteczność w ich przewidywaniu. Znacznie rzadziej mylił się jednak przy typowaniu zwycięstw wyjazdowych (48,3 % poprawnych typów, przy

45,9 % dla sieci neuronowej). Ogólnie można stwierdzić, że jakość obydwu modeli jest zbliżona, zatem prawdopodobnie oba algorytmy zostały poprawnie zastosowane. W dalszych analizach będę używał modelu zbudowanego algorytmem sieci neuronowych.

Rozdział 3 – Ocena modelu i porównanie do modeli wielkich firm bukmacherskich

W poprzednim rozdziale został oszacowany model, którego trafność na zbiorze testowym wyniosła 49,91 %. Wynik niemal połowy meczów Bundesligi od sezonu 2016/17 do marca 2018/19 został poprawnie przewidziany. Teraz należy porównać taki model z innymi możliwymi do stworzenia modelami w celu uzyskania punktu odniesienia i ostatecznej oceny jakości modelu.

Model sieci neuronowych a model prosty

Najprostszym modelem do stworzenia jaki przychodzi na myśl, jest model, w którym we wszystkich meczach przewidujemy wygraną drużyny domowej. W okresie analizowanym w zbiorze testowym udział meczów wygranych przez drużynę domową wynosi 46,66 %, zatem model zbudowany w ten sposób osiągnąłby trafność o ponad 3 pkt. procentowe gorszą.

Kolejnym punktem odniesienia może być model, w którym każdy mecz dwóch najlepszych drużyn w historii Bundesligi (Bayernu i Borussii) klasyfikujemy jako zwycięski (tzw. każdy mecz Bayernu oraz każdy mecz Borussii oprócz meczów przeciwko Bayernowi) a resztę meczów klasyfikujemy jako zwycięstwo drużyny domowej. W takim wypadku osiągnęlibyśmy wynik 50,64 %. Jest to zatem wynik lepszy niż wynik osiągany przez model wytrenowany za pomocą sieci neuronowych. Można uznać, że na trzech sezonach poddanych analizie bardziej opłacałoby się korzystać z takiego, prostego modelu niż typować wyniki za pomocą modelu zbudowanego przeze mnie.

Trafność modelu stworzonego a trafności modeli firm bukmacherskich

Przy ocenie jakości modelu warto dokonać porównań z modelami stworzonymi przez innych. Modele piłkarskie są modelami bardzo popularnymi i często wykorzystywanymi. Najlepsze modele należą do firm bukmacherskich, szczególnie do tych, które na rynku utrzymują się już od wielu lat. Uzyskane przeze mnie dane dotyczą kursów bukmacherskich ustalonych na mecze Bundesligi przez cztery firmy należące do największych i najbardziej znanych na świecie. Są to William Hill, Bet365, Bet&Win i Interwetten. Wszystkie te firmy utrzymują się na rynku od ponad 15 lat.

Na podstawie danych można wysnuć kilka ciekawych wniosków:

1. Firmą, która najlepiej przewidywała rezultat meczów piłkarskich, była firma Bet365. Średnia trafność predykcji na sezony 2004/2005 do sezonu 2018/2019 wyniosła 0,512.

Oznacza to, że ponad połowa rezultatów meczów na ten okres została poprawnie przewidziana. Jakość modeli pozostałych firm jest tylko nieco gorsza, najmniej dokładny model stworzyła firma William Hill o trafności 0,509. Różnice są zatem bardzo niewielkie.

2. Generalnie trafność obecnych predykcji jest wyższa niż trafność predykcji 10 lat temu. Wynika to prawdopodobnie z faktu uzyskiwania przez firmy coraz większej ilości danych dotyczących meczów, zarówno większej ilości zmiennych jak i dłuższego horyzontu czasowego.
3. Sezonami, które okazały się najbardziej przewidywalne, to znaczy jakość predykcji dokonanych na te sezony była najwyższa, były sezony 2013/2014[12] (firma Interwetten uzyskała na tym sezonie rekordowy wynik – 56,86 % właściwie przewidzianych wyników meczów) oraz sezon 2018/2019 do marca 2019. Co charakteryzowało te dwa sezony? Na czele końcowej tabeli znalazły się drużyny, które powszechnie uznaje się za faworytów. W sezonie 2013/2014 pierwsze cztery miejsca zajęły drużyny o najwyższej średniej ilości punktów w historii Bundesligi, czyli Bayern, Borussia, Bayer i Schalke. Można zatem uznać wyniki tego sezonu za przewidywalne i nie jest zaskoczeniem wysoka jakość predykcji na ten sezon. Tabela sezonu 2018/2019 pod koniec marca 2019 roku również wyglądała przewidywalnie. Na czele znalazły się Bayern, Borussia i RB Lipsk.
4. Sezonem najmniej przewidywalnym okazał się sezon 2010/2011[13]. Średnia jakość predykcji wszystkich czterech firm bukmacherskich wyniosła tylko nieco ponad 48 %. Tak kiepską jakość modeli można tłumaczyć na dwa sposoby. Po pierwsze sezon ten miał miejsce już 9 lat temu, kiedy ilość dostępnych danych była znacznie mniejsza niż teraz. Gdyby podobnie nieprzewidywalny sezon zdarzył się obecnie, prawdopodobnie jakość zbudowanych modeli byłaby znacznie wyższa. Po drugie rzuca się w oczy zaskakująco słaba postawa Bayernu Monachium, Schalke 04 oraz Wolfsburga a zaskakująco dobra postawa takich drużyn, jak Hannover i Mainz.

Jak natomiast przedstawiała się jakość modeli przetestowanych na danych z lat 2016/17 – 2018/19? Średnia jakość predykcji na tych sezonach wyniosła od 50,76 % dla firmy Interwetten do 51,93 % dla firmy Bet&Win. Trafności predykcji są zatem nieco wyższe od trafności mojego modelu, jednak jest to przewaga w wysokości 1-2 pkt. procentowe.

Jak wcześniej zostało ustalone trafność prostego modelu, który zawsze przewiduje wygraną Bayernu i Borussii, a następnie drużyny domowej, wyniosła w latach 2016/17 –

2018/19 około 50,64 %. Oznacza to, że modele firm bukmacherskich mają trafność niewiele lepszą od takiego niemal losowego obstawiania wyniku. Okazuje się, że korzystanie z modeli w bardzo niewielkim stopniu ułatwia obstawienie prawidłowego wyniku, a w niektórych przypadkach korzystanie z modelu jest całkowicie pozbawione sensu. Co może być przyczyną tak słabej trafności modeli? Po pierwsze, modele bukmacherskie nie służą do obstawiania prawidłowego wyniku meczu, lecz do wyliczenia prawdopodobieństwa danego wyniku. Załóżmy sytuację w której pierwsza drużyna zgodnie z modelem ma 41 % szans na zwycięstwo, na remis szansa wynosi 20 % a na zwycięstwo drugiej drużyny – 39 %. Jeżeli wygra drużyna druga, predykcja na ten mecz okaże się nieprawidłowa, natomiast z punktu widzenia firmy bukmacherskiej to, czy wygra drużyna domowa czy wyjazdowa nie ma praktycznie żadnego znaczenia. Po drugie, na wyniki meczów piłkarskich wpływa wiele czynników losowych lub czynników, które bardzo ciężko uwzględnić w modelu. Po za takimi oczywistymi czynnikami jak „dyspozycja dnia", czy fakt, że zawodnicy danej drużyny zabalowali w noc przed meczem w pobliskim hotelu, są to:

1. Skład i taktyka drużyny. Są one podawane zazwyczaj na około godzinę przed meczem, czego przyczyną jest fakt, że trener drużyny chce do końca utrzymać w niepewności trenera drużyny przeciwnej. Od składu drużyny może zależeć bardzo wiele, a trener często przed ważniejszymi meczami wystawia skład rezerwowy, który ma znacznie mniejsze szanse na korzystny rezultat.
2. Kontuzje kluczowych zawodników. Przykładowo brak jednego z kluczowych piłkarzy może znacznie zmniejszyć szansę drużyny na zwycięstwo. Na przykład zgodnie z magazynem „Kicker", z Marco Reusem w składzie Borussia Dortmund zdobywa średnio 2,42 punkta na mecz, natomiast bez Reusa zaledwie 1,69 punkta na mecz. Jest to zatem różnica znacząca, a ciężka do uwzględnienia w modelu.

Jak bukmacherzy ustalają kursy?

Jednym z zadań zbudowanego modelu jest umożliwienie wyznaczenia jak najbardziej wiarygodnych kursów bukmacherskich na przyszłe mecze Bundesligi. Celem każdego bukmachera jest takie ustalenie kursów, aby zmaksymalizować długoterminowy zysk z prowadzonej działalności. Na podstawie znajomości prawdopodobieństwa poszczególnych zdarzeń, bukmacher może osiągnąć długoterminowy zysk niezależnie od wyniku pojedynczego meczu. Podstawowym problemem jest właśnie znajomość prawdopodobieństw, które na podstawie historycznych meczów są wyznaczane przez model.

Kursy w tej pracy zostały ustalone przy założeniu tak zwanej „sytuacji idealnej”. W celu wytłumaczenia tego pojęcia i zilustrowania metody wyznaczania kursów posłużę się przykładem: weźmy mecz Bayern Monachium vs Borussia Dortmund z 26 lutego 2011. Przybliżone prawdopodobieństwa poszczególnych zdarzeń według bukmachera podane są w tabeli 3.1 [14].

Tabela 3.1 Prawdopodobieństwa poszczególnych wyników meczu Bayern vs Borussia

Wygrana drużyny domowej (Bayern Monachium)	Remis	Wygrana drużyny wyjazdowej (Borussia Dortmund
50 %	30%	20%

Na podstawie tych prawdopodobieństw bukmacher ustalił kursy na poziomie umożliwiającym osiągnięcie pewnej marży. Załóżmy, że bukmacher chciałby osiągnąć zysk w wysokości 10 % z każdych otrzymanych 100 zł. Przykładowe wyznaczenie kursów przez bukmachera mogło wyglądać następująco: bukmacher wyznaczył prawdopodobieństwa danego rezultatu meczu. Następnie skorzystał ze wzoru (1-y)/p, gdzie y oznacza zakładaną marżę, natomiast p oznacza prawdopodobieństwo. W ten sposób uzyskał kursy ukazane w tabeli 3.2.

Tabela 3.2 Kursy na poszczególne wyniki meczu Bayern vs Borussia

Wygrana drużyny domowej (Bayern Monachium)	Remis	Wygrana drużyny wyjazdowej (Borussia Dortmund
(100% - 10%)/50% = 1,8	(100% - 10%)/30% = 3	(100%-10%)/20% = 4,5

Bukmacher ustalił kursy: 1,8 na wygraną Bayernu, 4,5 na wygraną Borussii, 3 na remis. Oznacza to przykładowo, że w przypadku postawienia 10 zł na wygraną Bayernu obstawiający zakład może wygrać 10*1,8 = 18 zł.

W celu obliczenia faktycznego zysku posłużę się wspomnianą wcześniej „sytuacją idealną”. Załóżmy, że udział postawionych pieniędzy na każdy z trzech zakładów jest równy wielkości prawdopodobieństw tych wyników. Załóżmy zatem, że postawione zostało 50 zł na wygraną Bayernu, 30 zł na remis i 20 zł na wygraną Borussii. W tej sytuacji Bukmacher zainkasował 100 zł, lecz musi teraz wypłacić wynagrodzenia za dobrze obstawione zakłady:

- Przy wygranej Bayernu bukmacher musi wypłacić 50*1,8 = 90 zł
- Przy wygranej Borussii bukmacher musi wypłacić 20*4,5 = 90 zł
- Przy remisie bukmacher musi wypłacić 30*3 = 90 zł

W ten sposób niezależnie od wyniku meczu bukmacher zarabia 10 zł. Oczywiście sytuacja taka jest mocno wyidealizowana, ze względu na fakt, że w pojedynczym meczu niekoniecznie postawione na każdy mecz pieniądze muszą być równe prawdopodobieństwom zdarzeń. Jednak w przypadku wielu zakładów i dużej liczby obstawiających osób, bukmacher w długim okresie powinien zainkasować zysk w wysokości 10%. Kluczowe jest jednak w tym wypadku właściwe wyznaczenie prawdopodobieństw, co jest głównym zadaniem modelu bukmacherskiego.

Powyżej przedstawioną metodologia zastosowana została również w tej pracy w celu wyliczenia kursów bukmacherskich na podstawie uzyskanych prawdopodobieństw oraz umożliwienia porównań kursowych pomiędzy modelami.

Porównania kursowe między modelami

Jak już zostało ustalone, prawdziwym celem bukmacherów jest wyznaczenie prawidłowych kursów opierających się na przewidywanym prawdopodobieństwie danego zdarzenia. Prawidłowych, czyli takich, które niezależnie od ostatecznego wyniku meczu zmaksymalizują zyski firmy.

Przy założeniu, że bukmacherzy ustalali kursy opisaną w poprzednim podrozdziale metodą, ich marża wynosi około 5-6 %. Przyjęta przeze mnie marża w wysokości 6% umożliwia zatem przybliżone porównania kursów.

Gdy spogląda się na kursy ustalane przez czterech analizowanych bukmacherów, są one do siebie bardzo zbliżone. Kursy ustalone przez stworzony przeze mnie model w niektórych przypadkach różnią się natomiast w znacznym stopniu od kursów ustalonych przez bukmacherów. Różnice szczególnie widoczne są w następujących aspektach:

1. Miejsce rozgrywania meczu: średni kurs firm bukmacherskich na drużyny grające u siebie wynosił 2,64 – 2,72. Średni kurs na drużyny domowe wynikający z mojego modelu wynosił około 2,34. Oznacza to znaczne przeszacowanie przez model znaczenia gry na własnym stadionie. Średni kurs na remis i zwycięstwo drużyny wyjazdowej znacząco się nie różni.
2. Niedocenianie siły drużyn, takich jak Bayern i Borussia: fakt ten jest szczególnie widoczny w meczach wyjazdowych. Średni kurs na wyjazdowe zwycięstwo Bayernu

i Borussii oszacowany przez bukmacherów wynosił 1,74 – 1,77. Średni kurs według mojego modelu wynosił około 1,94.

3. Przecenianie siły beniaminków. Średni kurs bukmacherów na zwycięstwo beniaminka wynosił 3,92 – 4,12, według mojego modelu było to 3,64.
4. Problem z obserwacjami nietypowymi. Najwyższe odchylenia kursów mojego modelu od kursów bukmacherskich odnotowano dla drużyny RB Lipsk w sezonie 2016/2017, kiedy drużyna jako beniaminek zdobyła wicemistrzostwo Bundesligi. Średnie kursy bukmacherów na zespół Lipska wynosiły od 2,39 do 2,44 (zaskakująco niska rozbieżność, jak na tak nietypowe obserwacje), natomiast kurs wynikający z mojego modelu wyniósł aż 3,80. Oznacza to znaczne niedocenienie drużyny. Problem mojego modelu z obsługą obserwacji dotyczących Lipska miał miejsce również w drugim sezonie (kursy bukmacherskie 2.31-2.41, średni kurs mojego modelu - 2.97). Co ciekawe natomiast, nietypowe obserwacje dotyczące bardzo kiepskiej postawy Schalke w sezonie 2018/2019 nie spowodowały różnic pomiędzy kursami.

Podsumowanie

Piłka nożna jest sportem bardzo trudnym do modelowania. Ze względu na duży wpływ czynników losowych („dyspozycja dnia", czy przypadkowa bramka strzelona w ostatniej minucie meczu) oraz czynników trudnych do uwzględnienia (kontuzja kluczowego piłkarza, skład i taktyka drużyny), modele przewidujące wynik meczu charakteryzują się niską użytecznością. Celem modelu bukmacherskiego nie jest jednak przewidzenie wyniku meczu, lecz wyznaczenie jak najbardziej trafnych prawdopodobieństw danego rezultatu. Na podstawie takich prawdopodobieństw bukmacher może wyznaczyć kursy, które przyniosą mu zysk niezależnie od końcowego wyniku meczu.

- Z analiz wynika, że bardzo istotny wpływ na wynik meczu Bundesligi ma miejsce jego rozgrywania. Niemal 47 % meczów zostało wygranych przez drużynę grającą na własnym stadionie. Tylko 25 % stanowiły remisy.
- Ostateczny zbiór danych użyty do modelowania dotyczył meczów Bundesligi od sezonu 1996/97 do marca sezonu 2018/19. Zmienną objaśnianą w modelu była zmienna dotycząca rezultatu spotkania (wygrana drużyny domowej, wygrana drużyny wyjazdowej, remis). Zmienne objaśniające bazowały na liczbie punktów i dotyczyły ogólnej jakości drużyny (średnia liczba punktów w trzech ostatnich sezonach), formy drużyny (średnia liczba punktów w pięciu ostatnich meczach), formy drużyny w meczach u siebie i na wyjeździe (średnia liczba punktów w pięciu ostatnich meczach u siebie i na wyjeździe) oraz liczby punktów w ostatnich czterech meczach pomiędzy konkretnymi dwoma drużynami.
- Model został wytrenowany na danych z sezonów 1996/97 do 2015/16 i przetestowany na danych z sezonów 2016/17 do marca 2018/19. Do treningu zostały użyte dwa algorytmy: lasy losowe i sieci neuronowe.
- Jakość predykcji uzyskanej za pomocą lasów losowych na zbiorze treningowym wyniosła około 51 % natomiast na zbiorze testowym około 49,95 %.
- Jakość predykcji uzyskanej za pomocą sieci neuronowych na zbiorze treningowym wyniosła około 51,20 % natomiast na zbiorze testowym 49,91 %. Różnice w trafności modeli uzyskanych obiema metodami są niewielkie, zatem można uznać, że algorytmy zostały poprawnie wykorzystane.
- Porównania wykazały, że bardziej opłacalne od używania modelu w okresie 2016/17 – 2018/19 było proste typowanie za każdym razem zwycięstw Bayernu i Borussii

a w następnej kolejności drużyny grającej u siebie. W takim wypadku osiągnęłoby się 50,64 % trafności.

- Modele firm bukmacherskich okazały się niewiele lepsze w typowaniu wyniku meczu, niż model z poprzedniego punktu. Średnia jakość modeli firm Interwetten, Bet&Win, Bet365 i William Hill wynosiła około 50,76 % do 51,93 %. Najwyższą jakość osiągnął model firmy Bet&Win.
- Kursy bukmacherskie na większość historycznych meczów Bundesligi okazały się dla każdego z czterech bukmacherów bardzo zbliżone. Kursy wynikające ze stworzonego przeze mnie modelu w niektórych przypadkach mocno odstawały od kursów bukmacherów. Model przeszacowywał znaczenie gry u siebie, nie doceniał drużyn silnych takich jak Bayern i Borussia a przeceniał drużyny uznawane za słabsze, na przykład beniaminki. Dodatkowo kursy na obserwacjach nietypowych takich jak mecze Lipska z sezonu 2016/17 były znacznie wyższe niż kursy modeli bukmacherskich.

Okazuje się, że na podstawie samych historycznych punktów ciężko jest przewidzieć rezultat kolejnego spotkania a stworzony przeze mnie model nieco odstaje od modeli firm bukmacherskich. Usprawnienia, których można by dokonać to między innymi:

a) Zwiększenie liczby obserwacji poprzez dodanie danych o historycznych meczach innych lig (angielskiej, hiszpańskiej, włoskiej). Modele firm bukmacherskich z pewnością bazują na większej ilości obserwacji niż 7000 jak to było w moim przypadku.
b) Dodanie zmiennych dotyczących kontuzji kluczowych zawodników, liczby strzelonych bramek, oddanych strzałów i tym podobne. Dane takie są znacznie trudniej dostępne od danych dotyczących liczby punktów, jednak mogłyby okazać się dobrymi predyktorami rezultatu meczu.
c) Zastanowienie się nad lepszym uzupełnieniem braków danych. Prawdopodobnie jest to kluczowy problem stworzonego przeze mnie modelu, gdyż braki danych dotyczyły około 70 – 80 % obserwacji. Ich lepsze uzupełnienie poprawiłoby zapewne jakość modelu

Wygląda na to jednak, że przysłowie „piłka nożna jest piękna bo nieprzewidywalna" okazało się znaleźć potwierdzenie w przeprowadzonej analizie. I może niech tak zostanie, bo za to przecież właśnie ją kochamy.

[1]*https://www.przegladsportowy.pl/pilka-nozna/ligi-zagraniczne/laliga/real-madryt/sergio-ramos-pobil-niekorzystny-rekord-obejrzal-najwiecej-czerwonych-kartek/se46rjs*

[2] Podane przykłady są akurat wymyślone, ale dobrze oddają ogólny charakter ciekawostek i statystyk podawanych niekiedy przez komentatorów

[3]*https://www.businessinsider.com/the-20-most-popular-rich-list-football-teams-on-social-media-2018-8?IR=T#1-real-madrid-cf-2019-million-followers-20*

[4]*http://www.football-data.co.uk/germanym.php*

[5]Wyboru horyzontu trzyletniego dla określenia jakości drużyny i horyzontu pięciomeczowego dla określenia formy drużyny dokonałem na podstawie oglądanych meczów piłki nożnej i obserwacji empirycznych

[6]Prawidłowo przewidziany rezultat meczu to faktyczny rezultat (wygrana drużyny domowej, wygrana drużyny wyjazdowej, remis), który pokrywał się z przewidywaniami. Przy wyliczaniu trafności predykcji bukmacherskich brałem pod uwagę kurs na dane zdarzenie. Przykładowo: Kurs na mecz Werder Brema vs Schalke 04 z 6 sierpnia 2004 według Bet365 wynosił: 1,9 na zwycięstwo Werderu, 3,25 na remis i 3,5 na zwycięstwo Schalke. Najniższy kurs bukmacher wyliczył na zwycięstwo Werderu a zatem uznał, że zdarzenie to jest najbardziej prawdopodobne. Zgodnie z przewidywaniami Werder wygrał mecz. Predykcja okazała się zatem trafna i pokrywała się z faktycznym rezultatem meczu

[7]Beniaminek to drużyna, która awansowała do ligi w danym sezonie

[8]*http://pogotowiestatystyczne.pl/slowniczek/r-pearsona/*

[9]Na podstawie podręcznika M.Szeligi *Data Science i Uczenie Maszynow,* rozdział 5: Klasyfikacja, str. 128

[10]W tej pracy nie będę wchodził w szczegóły, omówienie sposobów podziału znajduje się tutaj: *https://datascience.stackexchange.com/questions/10228/gini-impurity-vs-entropy*

[11]Proces znajdowania optymalnych parametrów jest czasochłonny a algorytm ma do przeszukania w tym wypadku 3*3*2 = 18 zestawów parametrów. Każdy zestaw został sprawdzony 10 – krotną walidacją krzyżową, zatem model musiał zostać wytrenowany 180 razy, co łącznie zajęło około 3 godziny. Zatem ze względu na ograniczenia czasowe i technologiczne zdecydowałem się nie sprawdzać większej liczby parametrów.

[12]*https://pl.wikipedia.org/wiki/Bundesliga_niemiecka_w_pi%C5%82ce_no%C5%BCnej_(2013/2014)*

[13]*https://pl.wikipedia.org/wiki/Bundesliga_niemiecka_w_pi%C5%82ce_no%C5%BCnej_(2010/2011)*

[14] dane wymyślone

Źródła

[1] Szeliga M.: *Data Science i Uczenie Maszynowe*

[2] Zocca V., Spacanga G., Slater D. Roelants P.: *Deep Learning, Uczenie głębokie z językiem Python, Sztuczna inteligencja i sieci neuronowe*

[3] Kuper S., Szymański S.: *Futbonomia*

[4] Jeffries J., Oliver C.: *The book on bookies*

[5] Cortis, D.: *Expected Values and variance in bookmaker payouts: A Theoretical Approach towards setting limits on odds*

[6] https://datahub.io/sports-data/german-bundesliga

[7] https://en.wikipedia.org/wiki/Mathematics_of_bookmaking

[8] http://www.football data.co.uk/germanym.php

[9] https://pl.wikipedia.org/wiki/Bundesliga_niemiecka_w_pi%C5%82ce_no%C5%BCnej

[10] https://www.onlinebetting.org.uk/betting-guides/biggest-bookies.html

[11] https://pogotowiestatystyczne.pl/slowniczek/r-pearsona/

[12] https://cran.r-project.org/web/packages/ranger/ranger.pdf

[13] https://www.pinnacle.com/en/betting-articles/Betting-Strategy/Calculate-margins-on-1X2-odds/BHHJ5TXM8PFDHSYX