Czy na podstawie statystyk meczowych zawodnika możemy przewidzieć jego żółte kartki?

1. Wstęp:

Piłka nożna to dynamiczna i emocjonująca gra, w której każdy zawodnik odgrywa istotną rolę w osiągnięciu sukcesu przez swoją drużynę. Jednak oprócz zdolności technicznych i taktycznych, ważnym aspektem gry jest również dyscyplina i fair play na boisku. Jednym z czynników oceniających postawę zawodników jest liczba otrzymanych żółtych kartek.

Ich liczba w piłce nożnej stanowi istotny element oceny i monitorowania zachowań piłkarzy na boisku. Nie tylko wpływa ona na wynik meczu, ale również może mieć konsekwencje dla kolejnych spotkań zawodników, a nawet drużyny jako całości. W tym raporcie przeprowadzę analizę liczby żółtych kartek w zależności od indywidualnych statystyk piłkarzy, aby lepiej zrozumieć czynniki wpływające na ich tendencję do otrzymywania kar.

Celem tej analizy jest uzyskanie głębszego wglądu w związki między poszczególnymi aspektami gry piłkarzy a ich skłonnością do otrzymywania żółtych kartek. Przeanalizuję różne statystyki indywidualne, takie jak liczba fauli, przewinienia, agresywne zachowanie, a także statystyki związane z rolą piłkarza na boisku, takie jak pozycja, czas gry.

Wyniki tej analizy mogą mieć praktyczne zastosowanie dla trenerów, menedżerów drużyn i sędziów, którzy mogą użyć tych informacji do podejmowania decyzji dotyczących taktyki, selekcji zawodników i oceny ich zachowania na boisku. Ponadto, ta analiza może przyczynić się do zwiększenia świadomości wśród piłkarzy na temat ich własnego zachowania i konsekwencji, które mogą wynikać z otrzymania żółtych kartek.

W kolejnych sekcjach tego raportu przedstawię szczegółowe wyniki analizy oraz wnioski, które mogą być użyteczne dla zainteresowanych stron. Omówię również ewentualne ograniczenia naszej analizy oraz sugestie dla dalszych badań w tej dziedzinie.

W pracy przedstawię 2 główne próby sprawdzenia, czy liczba zdobytych żółtych kartek przez zawodników w sezonie 2021/2022 jest w jakikolwiek sposób zależna od ich indywidualnych statystyk. Aby to zbadać wykorzystam model regresji, aby móc jak najlepiej przewidzieć wartości dla wszystkich zawodników.

Zależność ta zostanie przeanalizowana dla zawodników z TOP 7 najlepszych lig europejskich oraz drugiego stopnia rozgrywkowego w Anglii. Wykorzystane ligi są następujące:

* angielska Premier League
* włoska Serie A
* niemiecka Bundesliga
* hiszpańska La Liga
* francuska Ligue 1
* portugalska Primeira Liga
* holenderska Eredivisie
* angielskie Championship

Powyższe ligi zostały wybrane ze względu na kilka czynników. Pierwszym z nich jest ich poziom sportowy. W tych ligach występują niektóre z najlepszych drużyn i piłkarzy na świecie. Zawodnicy grający w tych ligach są często uważani za elitę piłki nożnej, co czyni je interesującymi pod względem analizy ich indywidualnych statystyk. Kolejnym czynnikiem jest różnorodność stylów gry. Każda liga ma swoje charakterystyczne cechy i style gry. Liga angielska jest znana z intensywności i fizyczności, hiszpańska z techniki i umiejętności, a włoska z taktyki i defensywy. Analiza różnic między ligami może ujawnić potencjalne różnice w liczbie żółtych kartek, wynikające z tych różnic w stylach gry. Ostatnim czynnikiem jest popularność i zainteresowanie. Powyższe ligi cieszą się ogromnym zainteresowaniem zarówno lokalnie, jak i globalnie. Transmisje meczów z udziałem drużyn tych lig są obserwowane przez miliony kibiców na całym świecie. Analiza statystyk piłkarzy z tych lig może dostarczyć cennych informacji dla szerokiej publiczności.

1. Scrapowanie:

W celu zebrania danych statystycznych dotyczących piłkarzy do analizy, skorzystano z dwóch stron internetowych: fbref.com oraz soccerstats247.com. Proces zbierania danych został przeprowadzony przy użyciu bibliotek request oraz BeautifulSoup.

Szczegółowe i opisane kody znajdują się w plikach scraping\_fbref.py oraz scraping\_soccer\_stats.py.

Głównym celem tego kodu jest automatyczne scrapowanie1 danych piłkarzy z różnych lig na podstawie podanych adresów URL i zapisywanie ich w plikach CSV. Umożliwia również określenie różnych wzorców dla różnych typów statystyk, dzięki czemu można łatwo dostosować proces scrapowania do różnych rodzajów danych, takich jak statystyki standardowe lub różne statystyki dodatkowe.

Na stronie fbref.com dostępny jest szeroki zakres informacji dotyczących statystyk indywidualnych piłkarzy. Zbiór danych zawiera informacje takie jak narodowość, pozycja, wiek, liczba rozegranych meczów, gole, asysty, żółte i czerwone kartki, oraz wiele innych kluczowych wskaźników. Dane te były dostępne dla wielu lig z różnych krajów.

Kolejną stroną, z której pozyskano dane, było soccerstats247.com. Ta strona internetowa również dostarcza szeroką gamę statystyk piłkarskich, takich jak liczba rozegranych meczów, gole, asysty, żółte kartki i wiele innych. Dane te były dostępne dla wielu lig z różnych krajów, co pozwoliło na analizę statystyk piłkarzy na różnych poziomach rozgrywkowych.

Proces scrapowania danych z obu stron internetowych był realizowany zgodnie z zasadami etycznymi i prawnymi, a dane zostały zebrane w sposób zautomatyzowany i precyzyjny.

1. Zbiory danych i jego przetwarzanie:
   1. Wybór danych

Dane wykorzystane w analizie dotyczą zawodników z TOP 7 najlepszych lig europejskich oraz drugiego stopnia rozgrywkowego w Anglii. Wykorzystane ligi są następujące:

• angielska Premier League

• włoska Serie A

• niemiecka Bundesliga

• hiszpańska La Liga

• francuska Ligue 1

• portugalska Primeira Liga

• holenderska Eredivisie

• angielskie Championship

Powyższe ligi zostały wybrane ze względu na kilka czynników. Pierwszym z nich jest ich poziom sportowy. W tych ligach występują niektóre z najlepszych drużyn i piłkarzy na świecie. Zawodnicy grający w tych ligach są często uważani za elitę piłki nożnej, co czyni je interesującymi pod względem analizy ich indywidualnych statystyk. Kolejnym czynnikiem jest różnorodność stylów gry. Każda liga ma swoje charakterystyczne cechy i style gry. Liga angielska jest znana z intensywności i fizyczności, hiszpańska z techniki i umiejętności, a włoska z taktyki i defensywy. Analiza różnic między ligami może ujawnić potencjalne różnice w liczbie żółtych kartek, wynikające z tych różnic w stylach gry. Ostatnim czynnikiem jest popularność i zainteresowanie. Powyższe ligi cieszą się ogromnym zainteresowaniem zarówno lokalnie, jak i globalnie. Transmisje meczów z udziałem drużyn tych lig są obserwowane przez miliony kibiców na całym świecie. Analiza statystyk piłkarzy z tych lig może dostarczyć cennych informacji dla szerokiej publiczności

* 1. Zbiory danych

W pracy zostały wykorzystane 2 zbiory danych:

1. Zbiór 1 ”Standards stats” [1] pozyskany ze strony [fbref.com](https://fbref.com/en/). Zbiór danych zawiera informacje dotyczące piłkarzy, takie jak ich narodowość, pozycja, wiek na początku sezonu oraz różne dane związane z ich występami i osiągnięciami, takie jak liczba meczów rozegranych, gole, asysty, żółte i czerwone kartki. Dodatkowo, zawiera także dane dotyczące oczekiwanych wyników, takie jak spodziewane gole, gole asystowane, oraz informacje dotyczące progresji i wskaźniki na 90 minut.
2. Zbiór 2 ”Miscellaneous stats” [2] pozyskany również ze strony [fbref.com](https://fbref.com/en/). Zbiór danych zawiera wszechstronne informacje dotyczące piłkarzy, ich występów i wyników, takie jak żółte kartki, czerwone kartki, faule, dośrodkowania, przechwyty, wygrane pojedynki powietrzne itp. Jest to kompleksowy zbiór danych umożliwiający analizę indywidualnych statystyk piłkarzy i ich wpływu na liczbę żółtych kartek, które otrzymują.

W ramach scrapowania zebrano dane ze strony soccerstats.com, lecz okazały się one niewystarczająco obszerne do analizy i ostatecznie zostały one niewykorzystane.

Powyższe zbiory dla ułatwienia dalszej pracy zostały połączone w jeden zbiór, który później został podzielony na 2 pliki zawierające dane treningowe (statystyki z sezonu 2021/22 dla 8 lig) oraz dane testowe (statystyki z sezonu 2022/23 z angielskiej Premier League i włoskiej Serie A).

Dla danych ze strony fbref.com, nazwy nagłówków pozostały bez zmian, zapisane w formie skrótów i w ramach większej czytelności, w plikach glossary\_from\_fbref.txt oraz glossary\_from\_fbref\_miscellaneous.txt znajdują się rozwinięcia skrótów i ich krótkie opisy.

* 1. Wstępne przetwarzanie danych

Obróbka danych zawodników została przeniesiona poza głównego notebook’a i kod przetwarzający dane znajduje się w pliku fbref\_data\_processing.py, gdzie zostały wytłumaczone kolejne czynności przetwarzające dane.

Ze zbiorów zostały odrzuceni zawodnicy, którzy nie rozegrali przynajmniej 3 pełnych meczów (atrybut ‘90s’ ≥ 3.0), zawodnicy, którzy byli zgłoszeni do kadry meczowej, lecz nigdy nie rozegrali żadnych minut na boisku. Powodem tego jest to, aby wyniki były bardziej rzetelne, gdyż zawodnicy grający mało zazwyczaj wchodzą z ławki na ostatnie minuty meczu, gdy już sama dynamika gry jest zdecydowanie mniejsza i poszczególne statystyki mogą odbiegać od tych normalnych.

Ponadto kod łączy moje dwa zbiory - Standard i Miscellaneous Stats – w jeden wspólny zbiór, niezawierający zdublowanych kolumn atrybutów oraz statystyk, które z góry przyjąłem za niepotrzebne tj. gole i asysty oczekiwane itd. Usunięta kolumna ‘Matches’ nie dotyczy liczby rozegranych meczy przez zawodnika, lecz była to kolumna zawierająca link do meczy rozegranych w sezonie przez danego zawodnika.

Jako że pozycja zawodnika nie jest wartością liczbową, w drugiej części zamieniono to na podstawie poniższego mapowania:

* bramkarz (GK) - 1
* obrońca (DF) – 2
* pomocnik (MF) – 3
* atakujący (FW) – 4
* obrońca, pomocnik (DF,MF) – 5
* obrońca, atakujący (DF,FW) – 6
* atakujący, pomocnik (FW,MF) – 7

W samych plikach pozostały wszystkie dane, które znajdowały się w zebranych zbiorach.

1. Analiza eksploracyjna

4.1 Ogólne przestawienie danych

Po przetworzeniu moich danych treningowych (all\_players\_2021\_2022.csv) i testowych (test\_2022\_2023.csv), moje dane zawierają statystyki dotyczące kolejno 3425 i 858 zawodników. Z tego można wyliczyć, że w analizie stosunek danych treningowych do danych testowych wynosi 80:20. Poniżej przedstawiono fragmenty danych treningowych, jak i testowych (ze względu na dużą liczbę poszczególnych statystyk, ograniczono je do tych najważniejszych1).

1 w samych plikach csv znajdują się wszystkie zebrane dane, które przed rozpoczęciem obliczeń i trenowania modeli zostały pomniejszone.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Lp. | Player | Pos | Age | MP | MIn | 90s | Gls | Fls | Fld | Int | TklW | Recov |
| 0 | Thomas Müller | 3 | 31.0 | 32 | 2558.0 | 28.4 | 8.0 | 11.0 | 22.0 | 15.0 | 20.0 | 94.0 |
| 1 | Manuel Neuer | 1 | 35.0 | 28 | 2510.0 | 27.9 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 85.0 |
| 2 | Joshua Kimmich | 3 | 26.0 | 28 | 2474.0 | 27.5 | 3.0 | 20.0 | 46.0 | 52.0 | 20.0 | 218.0 |
| 3 | Dayot Upamecano | 2 | 22.0 | 28 | 2211.0 | 24.6 | 1.0 | 28.0 | 11.0 | 32.0 | 42.0 | 249.0 |
| 4 | Serge Gnabry | 6 | 26.0 | 34 | 2185.0 | 24.3 | 14.0 | 11.0 | 21.0 | 14.0 | 31.0 | 146.0 |

Tabela 1 1

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Lp. | Player | Pos | Age | MP | MIn | 90s | Gls | Fls | Fld | Int | TklW | Recov |
| 0 | Rodri | 3 | 26 | 36 | 2911.0 | 32.3 | 2.0 | 47 | 20 | 33 | 41 | 301.0 |
| 1 | Erling Haaland | 4 | 22 | 35 | 2769.0 | 30.8 | 36.0 | 31 | 24 | 2 | 37.0 | 50.0 |
| 2 | Kevin De Bruyne | 3 | 31 | 32 | 2417.0 | 26.9 | 7.0 | 20 | 29 | 12 | 97.0 | 50.0 |
| 3 | İlkay Gündoğan | 3 | 32 | 31 | 2353.0 | 26.1 | 8.0 | 22 | 25 | 18 | 155.0 | 61.3 |
| 4 | Manuel Akanji | 2 | 27 | 29 | 2287.0 | 25.4 | 0.0 | 24 | 12 | 25 | 136.0 | 40.4 |

Tabela 2 1



4.2 Analiza poszczególnych danych

Obraz zawierający tekst, diagram, zrzut ekranu, krąg

Opis wygenerowany automatycznie

Rysunek 1:

Powyższy diagram przedstawia liczbę piłkarzy grających na danej pozycji. Jak widać niektórzy zawodnicy grają jednocześnie na różnych pozycjach. Najwięcej w danych treningowych mamy informacji dotyczących obrońców (1087), zaś najmniej – obrońców/atakujących (88).

Obraz zawierający tekst, zrzut ekranu, Wielobarwność, diagram

Opis wygenerowany automatycznie

Rysunek 2:

Powyższy diagram przedstawia średnią liczbę żółtych kartek na 90 minut zdobytych przez zawodników na konkretnych pozycjach. Jak widać średnio na 90 minut najwięcej żółtych kartek otrzymują pomocnicy (0.26 kartki/90min), zaś najmniej – bramkarze (0.06 kartki/90min).

Obraz zawierający tekst, zrzut ekranu, Wykres, Wielobarwność

Opis wygenerowany automatycznie Obraz zawierający tekst, zrzut ekranu, Wielobarwność, Wykres

Opis wygenerowany automatycznie

Rysunek 3/4:

Powyższe porównanie przedstawia średnią liczbę żółtych kartek na 90 minut zdobytych przez zawodników na konkretnych pozycjach przy określonej liczbie rozegranych meczów. Można zauważyć, że zawodnicy, którzy rozegrali więcej meczy, zdobyli stosunkowo mniej żółtych kartek na rozegrane 90 minut. Być może jest to spowodowane większym rozłożeniem się zdobytych żółtych kartek w liczbie przegranych minut na boisku.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Lp. | **Player** | **Position** | **Club** | **YellowCardsPer90** |
| 1 | Béni Makouana | FW,MF | montpellier | 1.111111 |
| 2 | Nikola Kalinić | FW | hellas verona | 1.111111 |
| 3 | Aleksandar Sedlar | DF,MF | mallorca | 1.111111 |
| 4 | Kevin-Prince Boateng | FW,MF | hertha bsc | 1.038961 |
| 5 | David Tavares | MF | famalicao | 1.000000 |

Tabela 3:

Powyższa tabela przedstawia 5 zawodników z największą liczbą zdobywanych żółtych kartek na rozegrane 90 minut. ‘’Najlepsi” zawodnicy średnio zdobywali ponad 1 żółtą co każdy rozegrany pełny mecz.

| Lp. | **Player** | **Position** | **Club** | **FoulsPer90** |
| --- | --- | --- | --- | --- |
| 1 | Yan Brice Eteki | MF | granada | 4.736842 |
| 2 | Tom Eaves | FW | hull city | 4.214286 |
| 3 | Martin Adeline | FW,MF | reims | 4.102564 |
| 4 | Julio Romão | DF,MF | santa clara | 4.000000 |
| 5 | Monchu | MF | granada | 3.918919 |

Tabela 4:

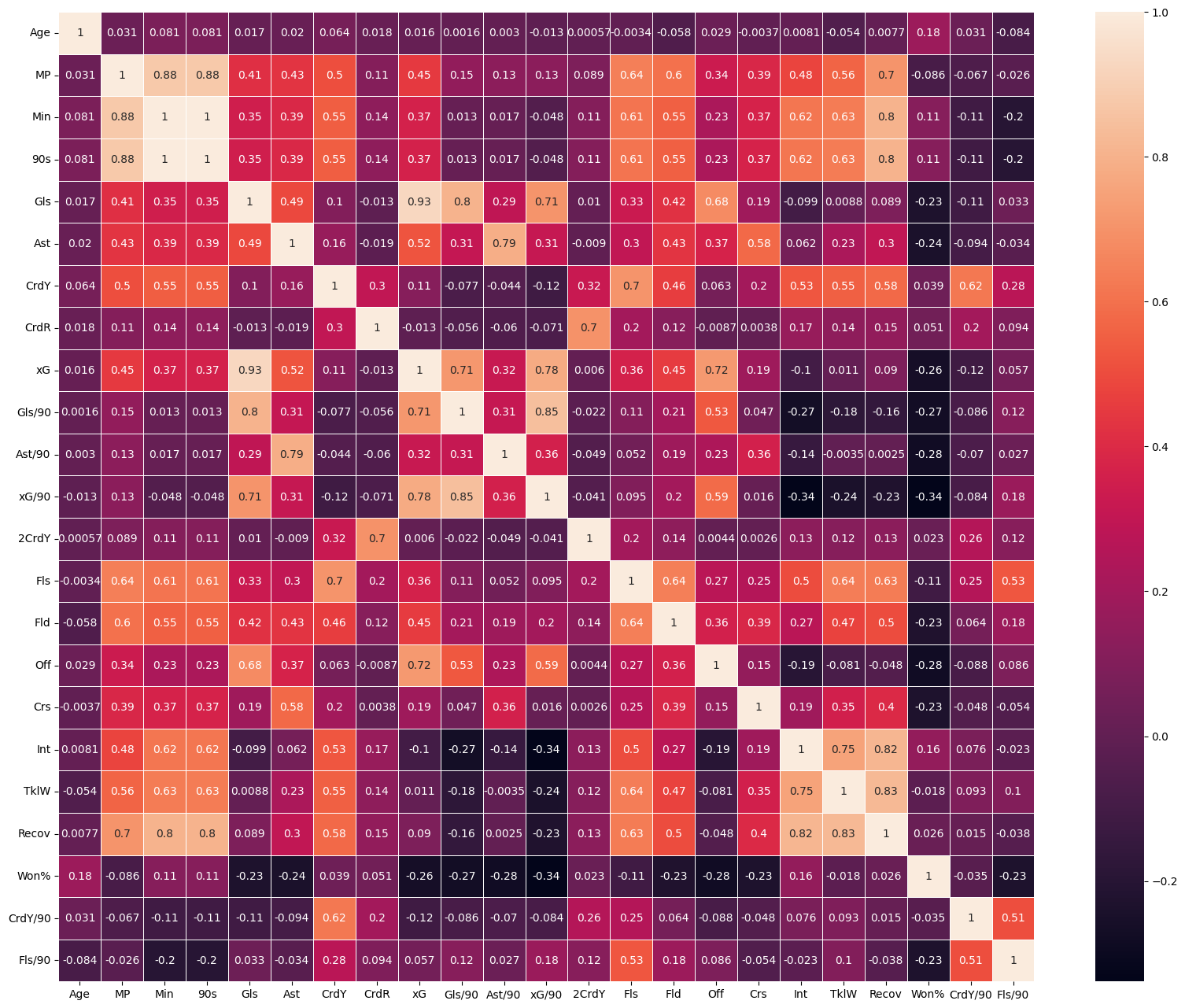
Powyższa tabela przedstawia 5 zawodników z największą liczbą popełnionych fauli na mecz. Co ciekawe, żaden z zawodnik nie widnieje w obydwóch tabelach.

Obraz zawierający linia, zrzut ekranu, diagram, Prostokąt

Opis wygenerowany automatycznie

Rysunek 5:

Powyższy diagram przedstawia średnią liczbę zdobytych żółtych kartek w sezonie przez zawodników pogrupowanych według wieku. Jak widać, słupek diagramu dla 16-latków znacznie przewyższa pozostałe. Jest to pewnie spowodowane małą liczbą zawodników w tym wieku i dużą liczbą kartek jednego zawodnika. Pionowe kreski oznaczają odchylenie standardowe dla danej grupy wiekowej i można zauważyć, że dla 16-letnich zawodników ta kreska jest praktycznie nie widoczna, co mogłoby potwierdzać moje wcześniejsze wytłumaczenie. Odchylenie standardowe dla najstarszych zawodników jest widocznie większe, co świadczy o dużej rozbieżność w kwestii liczby kartek.



Rysunek 6:

Ostatni w tej części diagram przedstawia macierz współczynników korelacji pomiędzy poszczególnymi statystykami piłkarzy. Im jaśniejszy jest kwadrat, tym większa dodatnia korelacja. Im ciemniejszy jest kwadrat, tym większa ujemna korelacja.

1. Trenowanie i testowanie modeli regresji
   1. Krótkie omówienie

Moje badania zostały podzielone na 2 części. Pierwsza część dotyczy przewidzenia liczby żółtych kartek zdobytych na 90 minut przez zawodników, zaś druga część skupia się na przewidzeniu liczby żółtych kartek w sezonie, na podstawie statystyk piłkarzy.

* 1. Modele

Aby maksymalnie wykorzystać potencjał, do moich obliczeń wykorzystałem następujące modele1:

* LinearRegression z biblioteki sklearn
* XBGRegressor z biblioteki xgboost
* LGBMRegressor z biblioteki lightgbm
* Neural Newtorks z biblioteki keras
* DecisionTreeRegressor z biblioteki sklearn

5.2.1 Model LinearRegression jest podstawowym modelem regresji liniowej, który stara się znaleźć najlepsze dopasowanie liniowe między zależną zmienną a zestawem niezależnych zmiennych. Wykorzystuje metodę najmniejszych kwadratów, aby minimalizować różnicę między rzeczywistymi a przewidywanymi wartościami.

5.2.2 Model XGBRegressor to implementacja algorytmu gradient boosting zespołowego (gradient boosting machine) opartego na drzewach decyzyjnych. Wykorzystuje technikę wzmacniania gradientowego, gdzie iteracyjnie dodaje drzewa decyzyjne, minimalizując gradient funkcji straty. Ten model jest znany ze swojej wysokiej wydajności i zdolności do radzenia sobie z dużymi i złożonymi zbiorami danych.

5.2.3 Model LGBMRegressor to kolejna implementacja gradient boosting zespołowego, która wykorzystuje drzewa decyzyjne z optymalizacją na podstawie histogramu. Działa na podobnej zasadzie jak XGBRegressor, ale zastosowana technika optymalizacji pozwala na szybsze trenowanie modelu i wykorzystanie mniej zasobów obliczeniowych.

5.2.4 Modele sieci neuronowych pozwalają na tworzenie różnych architektur sieci neuronowych, które są zdolne do uczenia się nieliniowych zależności i są często stosowane w zadaniach regresji, gdzie zależności są bardziej skomplikowane.

5.2.5 Model DecisionTreeRegressor to model regresji oparty na drzewach decyzyjnych, które rozdzielają przestrzeń cech na podstawie warunków, aby przewidywać wartość zmiennej zależnej. Drzewa decyzyjne podziałem zbioru danych na podstawie różnych cech, tworząc hierarchiczne struktury decyzyjne.

Podczas tworzenia wybranych modeli zastosowano metodę GridSearch do optymalizacji hiperparametrów.

1 więcej informacji dotyczących modeli i indywidualnych parametrów użytych w obliczeniach zostały przedstawione w notebook’u z obliczeniami notebook.py

5.3 Cechy

W pierwszej części, aby dokonać predykcji liczby żółtych kartek na 90 minut wykorzystano następujące cechy piłkarzy:

* wiek
* liczba fauli na 90 minut
* liczba rozegranych 90 minut
* liczba goli na 90 minut
* liczba wślizgów
* liczba przechwytów

W drugiej części do przewidzenia liczby żółtych kartek w sezonie użyto poniższych statystyk:

* pozycja zawodnika
* całkowita liczba rozegranych minut w sezonie
* całkowita liczba fauli w sezonie
* całkowita liczba wymuszonych fauli na przeciwniku w sezonie
* całkowita liczba wślizgów w sezonie
* całkowita liczba przechwytów w sezonie
* całkowita liczba odzyskanych piłek

5.4 Analiza obliczeń i wyników

Zacznijmy omówienie pierwszej części analizy, w ramach której na początku dokonanowyboru najlepszego stopnia wielomianu dla moich modelów regresji na najprostszym modelu regresji – LinearRegression. Do przetestowania modelu i sprawdzenia r2  wykorzystałem zarówno dane testowe, jaki i dane treningowe.

Oto wyniki:

|  |  |  |
| --- | --- | --- |
| stopień wielomianu | treningowe r2 | testowe r2 |
| 2 | 0.3066558083685973 | 0.3012091880918263 |
| 3 | 0.3224737780596829 | 0.294227351285141 |
| 4 | 0.35217616494387294 | 0.28816879166763076 |
| 5 | 0.4025056564697467 | 0.1560022347924197 |
| 6 | 0.4491209927724539 | -16.687128714386596 |
| 7 | 0.5809277033322346 | -505.44356265185615 |
| 8 | 0.714423821363342 | -101212.57251351075 |

Jak widać im wyższym stopień wielomianu, tym lepsze dopasowanie modelu do wykorzystanych danych treningowych, lecz wraz ze wzrostem, drastycznie maleje r2 dla danych testowych, co nie jest dobrą oznaką. Z tego też powodu została podjęta decyzja, aby wykorzystać domyślny wielomian - stopnia drugiego - gdyż daje porównywalne wyniki dla danych treningowych i testowych, co może pomóc znaleźć odpowiedni model oraz rezultat.

Poniżej przedstawiono wyniki wykorzystanych w obliczeniach modeli

Szczegółowe wyniki dla LinearRegression:

Wielomian

$y = 0.00301918x1 + 0.12637597x2 - 0.00132616x3 - 0.11237664x4 - 0.00019764x5 + 0.00118643x6 - 0.00259398$, gdzie:

y – przewidywana liczba żółtych kartek na 90 minut

x1 – wiek piłkarza (‘Age’)

x2 – liczba fauli na 90 minut (‘Fls/90’)

x3 – liczba rozegranych 90 minut (‘90s’)

x4 – liczba goli na 90 minut (‘Gls/90’)

x5 – liczba wślizgów (‘TklW’)

x6 – liczba przechwytów (‘Int’)

W ramach wyjaśnienia moich wyników, poniżej przedstawiono statystyki które zostały obliczone w celu określenia jakości rezultatów modelu.

r2 (ang. r-squared) jest miarą jakości dopasowania modelu regresji do danych. Oznacza ono współczynnik determinacji i jest liczbową wartością między 0 a 1. Wyższa wartość r2 wskazuje na lepsze dopasowanie modelu do danych.

MAE (ang. mean absolute error) to miara błędu średniego bezwzględnego w modelach regresji. Mierzy ona przeciętną różnicę między prognozowanymi wartościami modelu a rzeczywistymi wartościami danych.

MAE oblicza się poprzez zsumowanie wartości bezwzględnych różnic między prognozami a rzeczywistymi wartościami, a następnie podzielenie przez liczbę przykładów danych. Jest to miara bezwzględna, co oznacza, że nie bierze pod uwagę kierunku ani znaku błędu.

MSE (ang. mean squared error) to miara błędu średniokwadratowego w modelach regresji. Jest to jedna z najczęściej stosowanych miar błędu, która mierzy średnią kwadratową różnicę między prognozowanymi wartościami modelu a rzeczywistymi wartościami danych.

MSE oblicza się poprzez zsumowanie kwadratów różnic między prognozami a rzeczywistymi wartościami, a następnie podzielenie przez liczbę przykładów danych. Kwadratowa natura miary powoduje, że większe błędy mają większy wpływ na wynik niż mniejsze błędy.

Poniżej przedstawiono wyniki obliczeń predykcji liczby żółtych kartek na 90 minut1:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Rodzaj danych | r2 | MAE | MSE |
| LinearRegression | treningowe | 0.29130 | 0.10077 | 0.01834 |
| testowe | 0.29090 | 0.09634 | 0.01625 |
| XGBRegressor | treningowe |  |  |  |
| testowe |  |  |  |
| LGBMRegressor 1 | treningowe |  |  |  |
| testowe |  |  |  |
| LGBMRegressor 2 | treningowe |  |  |  |
| testowe |  |  |  |
| NN Keras 1 | treningowe |  |  |  |
| testowe |  |  |  |
| NN Keras 2 | treningowe |  |  |  |
| testowe |  |  |  |
| DecisionTreeRegressor 1 | treningowe |  |  |  |
| testowe |  |  |  |
| DecisionTreeRegressor 2 | treningowe |  |  |  |
| testowe |  |  |  |

1 wyniki zostały podane z dokładnością do 5 miejsc po przecinku.

Obraz zawierający tekst, mapa, zrzut ekranu, diagram

Opis wygenerowany automatycznieObraz zawierający tekst, diagram, mapa, zrzut ekranu

Opis wygenerowany automatycznie

Ze względu na to, że niemożliwe jest zwizualizowanie sensownego wykresu dla więcej niż 3 zmiennych, dla każdego wykorzystanego modelu, przedstawiono wykresy porównujące wyniki przewidziane przez modele z tymi prawdziwymi.

Przechodząc do drugiej części analizy, w tej części dokonano prób przewidzenia całkowitej liczby zdobytych żółtych kartek przez zawodników ciągu całego sezonu.

Poniżej przedstawiono wyniki wykorzystanych w obliczeniach modeli.

Szczegółowe wyniki dla LinearRegression:

Wielomian

y = -0.13721626x1 + 0.00021294 x2 - 0.11574336 x3 + 0.00635061 x4 - 0.03066969 x5 -0.00127684 x6 + 0.00028151x7 + 0.6805319, gdzie:

y – przewidywana liczba żółtych kartek w sezonie

x1 – pozycja piłkarza (‘Pos’)

x2 – całkowita liczba rozegranych minut w sezonie (‘Min’)

x3 – całkowita liczba fauli w sezonie (‘Fls’)

x4 – całkowita liczba wymuszonych fauli na przeciwniku w sezonie (‘Fld’)

x5 – całkowita liczba wślizgów w sezonie (‘Int’)

x6 – całkowita liczba przechwytów w sezonie (‘TklW’)

x7 – całkowita liczba odzyskanych piłek (‘Recov’)

Poniżej przedstawiono wyniki obliczeń predykcji liczby żółtych kartek w sezonie1:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Rodzaj danych | r2 | MAE | MSE |
| LinearRegression | treningowe |  |  |  |
| testowe |  |  |  |
| XGBRegressor | treningowe |  |  |  |
| testowe |  |  |  |
| LGBMRegressor 1 | treningowe |  |  |  |
| testowe |  |  |  |
| LGBMRegressor 2 | treningowe |  |  |  |
| testowe |  |  |  |
| NN Keras 1 | treningowe |  |  |  |
| testowe |  |  |  |
| NN Keras 2 | treningowe |  |  |  |
| testowe |  |  |  |
| DecisionTreeRegressor 1 | treningowe |  |  |  |
| testowe |  |  |  |
| DecisionTreeRegressor 2 | treningowe |  |  |  |
| testowe |  |  |  |

1 wyniki zostały podane z dokładnością do 5 miejsc po przecinku.

Obraz zawierający tekst, zrzut ekranu, linia, diagram

Opis wygenerowany automatycznieObraz zawierający tekst, zrzut ekranu, diagram, linia

Opis wygenerowany automatycznie

Porównanie badań i ich rezultatów

Badania wykazały, że żółte kartki są nie do końca przewidywalne i trudno jest je przewidzieć.

W pierwszej części dokonując predykcji liczby żółtych kartek na 90 minut, modele nie poradziły sobie najlepiej.

Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

Opis wygenerowany automatycznieObraz zawierający tekst, zrzut ekranu, diagram, Wykres

Opis wygenerowany automatycznie

Jak widać powyżej, mimo bardzo dobrego dopasowania się modeli XGBRegressor (0.882) oraz domyślnego DecisionTreeRegressor (aż 1.0) do danych treningowych, tak dla danych testowych okazały się najgorsze – DecisionTreeRegressor uzyskał nawet ujemny współczynnik r2  na poziomie -0.513, co oznacza praktycznie brak dopasowania.

Pozostałe modele dla danych treningowych poradziły sobie nie najlepiej. Co jest warte uwagi, to to, że model klasycznej regresji liniowej – LinearRegression – poradził sobie tak samo zarówno dla danych treningowych, jak i testowych. Ponadto model sieci neuronowych z warunkiem stopu mimo, że dla danych treningowych poradził sobie gorzej od klasycznego, to dla danych testowych poradził sobie nieznacznie lepiej.

Jak wiadomo, mimo że r2 pokazuje poziomo dopasowania wartości, tak nie jest jedyną miarą jakości modeli. Przejdźmy zatem do tego, jak wyglądają błędy modeli – MAE i MSE.

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznieObraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznieObraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

W tym przypadku MSE nie jest zbyt dobrym wyznacznikiem jakości, ze względu na to, że działamy w większości przypadkach na wartościach ułamkowych z przedziału od 0 do 1.

Skupmy się zatem na MAE i jak widać mimo błąd na poziomie nie wydaje się zbyt duży, to w porównaniu do mediany (dla danych treningowych - , dla danych testowych - ) pośród wszystkich liczby żółtych kartek zdobytych na 90 minut jest to duża wartość. Najmniejszy błąd procentowy dla danych testowych jest dla modelu klasycznej regresji liniowej na poziomie 50.53% (MAE = 0.09), zaś największy dla domyślnego DecisionTreeRegressor – 71.96% (MAE = 0.13). W przypadku gdyby piłkarz rozegrał niewiele meczy, modele zbytnio nie odbiegałyby od rzeczywistych wyników, lecz piłkarze w samej lidzie grają nawet po 40 meczów i w tym przypadku błąd byłby już znaczny – przynajmniej 3.6 kartki na 90 minut.

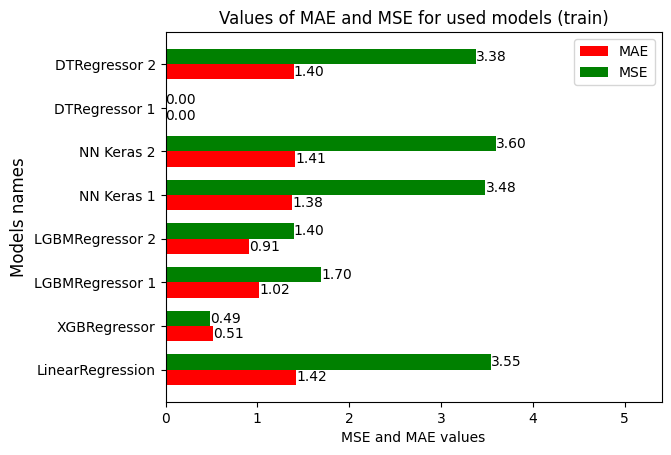
W drugiej części dokonując predykcji liczby żółtych kartek w całym sezonie, modele poprawiły swoje wyniki, lecz nie w takim stopniu, jak to było oczekiwane.

Obraz zawierający tekst, zrzut ekranu, diagram, Wielobarwność

Opis wygenerowany automatycznieObraz zawierający tekst, zrzut ekranu, Wielobarwność, diagram

Opis wygenerowany automatycznie

Podobnie jak w poprzedniej części, modele XGBRegressor i domyślny DecisionTreeRegressor poradziły sobie najlepiej, odnotowując wartości r2 na poziomach kolejno 0.938 oraz 1.0. Pozostałe modele też znacznie poprawiły swoje wyniki – największy wzrost w porównaniu do poprzedniej części dla danych testowych odnotowały: model XGBRegression (wzrost z poziomu 0.147 do 0.474), model sparametryzowanego DecisionTreeRegression (wzrost z poziomu 0.23 do 0.541) oraz model domyślnego DecisionTreeRegression (wzrost z -0.513 do -0.018), jednak w tym przypadku nie możemy mówić o poprawie modelu, gdyż dalej jest to wartość ujemna, zdecydowanie odbiegająca od oczekiwań.

 Obraz zawierający tekst, zrzut ekranu, wyświetlacz, numer

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, numer, wyświetlacz

Opis wygenerowany automatycznie Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

Podobnie jak to było wcześniej MSE nie będzie dobrym wyznacznikiem jakości wyników wytrenowanych modeli.

Jak widać najmniejszy błąd dla danych testowych generują modele sieci neuronowych i klasycznej regresji liniowej na poziomie 1.32/1.33 kartki, co moim zdaniem nie jest najgorszym wynikiem. Zważając też na to że liczba żółtych kartek jest liczbą całkowitą, to zaokrąglając błąd, wynik błędu na poziomie 1 jest całkiem dobrym wynikiem.

Porównując to do mediany żółtych kartek wśród zawodników równej 3 kartek, można stwierdzić, że mimo wszystko modele mogłyby lepiej to przewidzieć, gdyż błąd wynosi przynajmniej 40% wartości mediany i może wydawać się duży.

Sprawdźmy jeszcze, jak wygląda błąd po zaokrągleniu do najbliższej liczby całkowitej.

Obraz zawierający tekst, zrzut ekranu, numer, wyświetlacz

Opis wygenerowany automatycznie Obraz zawierający tekst, zrzut ekranu, wyświetlacz, numer

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, numer, wyświetlacz

Opis wygenerowany automatycznie Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

Jak widać statystyki nieznacznie się poprawiły, lecz zaokrąglanie i tak nie miało wielkiego wpływu na rezultaty. Najmniejszy błąd dla danych testowych po zaokrągleniu dalej mają modele sieci neuronowych oraz klasycznej regresji liniowej, kolejno 1.29 oraz 1.31. Poziom procentowy błędu w porównaniu do mediany praktycznie pozostaje bez zmian.

Wnioski

Podsumowując obliczenia i analizę można stwierdzić, że trudno jest przewidzieć współczynnik żółtych kartek do rozegranych 90 minut. Być może jest to spowodowane zbyt dużym rozrzutem i brakiem dużej korelacji wśród statystyk na podstawie dokonano predykcji lub złym dobraniem cech. Być może wybór innych lub większej ilości statystyk mogłoby poprawić wyniki tej predykcji. Także wyniki pokazały, że długie trenowanie modelu nie pomaga i wyniki danych testowych dla takich modeli są bardzo niesatysfakcjonujące. Za przykład z pewnością można wziąć domyślny model DecisionTreeRegression, który dla danych treningowych osiągał wynik r2 na perfekcyjnym poziomie 1.0, zaś dla danych testowych, jako jedyny model, osiągał r2 poniżej zera, co świadczy o kompletnym braku dopasowania się do danych.

Dla predykcji całkowitej liczby żółtych kartek w całym sezonie, modele poradziły sobie w miarę dobrze, myląc się średnio o ponad jedną kartkę na cały sezon. Można też stwierdzić, że wybór cech do tej predykcji okazał się bardziej trafny i wykorzystując je w poprzedniej predykcji, można otrzymać lepsze wyniki. Mimo że wyniki drugiej części ukazały się zadowalające, być może użycie modeli klasyfikacji, mogłoby jeszcze poprawić wyniki, przypisując każdej wartości liczby kartek, osobną klasę.

Z pewnością dalsza analiza i badania tego problemu mają ogromny potencjał. Taka wiedza może być przydatna dla sztabu szkoleniowego. Przewidywanie liczby żółtych kartek może dostarczyć informacji o tendencjach i stylu gry przeciwników oraz pomóc w opracowaniu planów gry i strategii dla własnej drużyny. Przy zastosowaniu innych metod i strategii obliczeń, wiedza dotycząca żółtych kartek może być niezwykle cenna w świecie piłki nożnej i warto poświęcić temu tematowi więcej uwagi.