AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Zarządzania

Samodzielna Pracownia Zastosowań Matematyki w Ekonomii

Praca dyplomowa licencjacka

Pozasportowe czynniki, wpływające na wynik meczu piłki nożnej

Nonsports factors determining the result of the football event

Autor:

Kierunek studiów: Informatyka i Ekonometria

Opiekun pracy:

Kraków, czerwiec 2020

Spis treści

Streszczenie

W pracy podjęto temat modelowania rezultatów meczów piłki nożnej, wśród których zaszło podejrzenie nadużycia, związanego z próbą wpłynięcia na wynik spotkania. W pierwszej części, przedstawiono wprowadzenie do tematu, jak i również teoretyczny wstęp, dotyczący użytych metod. Zbiór obserwacji bazował głównie na wielkości wahań kursowych w ofertach podmiotów bukmacherskich. Modele stworzono za pomocą dwóch technik: analizy dyskryminacyjnej oraz regresji logistycznej. W kolejnych fazach badania, poza próbą estymacji modelu optymalnego pod względem odsetka trafności reklasyfikacji, podjęto próbę wykorzystania modeli do predykcji prawdopodobieństw, detekcji podejrzanych spotkań. Stworzony został również prosty algorytm wspomagający decyzje inwestycyjne, w kontekście zawierania zakładów, jak i również tworzenia oferty bukmacherskiej.

1. Wprowadzenie

Możliwość zawierania zakładów na popularne wydarzenia sportowe w dzisiejszych czasach sprawiła, że ten rodzaj rozrywki stał się bardzo powszechny. Niektórzy nie wyobrażają sobie oglądania rozgrywek ulubionych drużyn bez emocjonującego dodatku w postaci hazardu. Poza najpopularniejszymi wydarzeniami, podmioty bukmacherskie starają się poszerzać swoją ofertę w jak największym stopniu. Spowodowane jest to chęcią osiągnięcia, jak największych zysków. Wzbogacanie oferty o nowe sporty, dodawanie możliwości zawierania zakładów w rozgrywkach bardziej niszowych to czynniki, które z pewnością pomagają przebić się na tle konkurencji, a co za tym idzie zwiększyć generowane dochody. Kierunek rozwoju zakładów bukmacherskich w takim kierunku stwarza również coraz większe możliwości manipulacji rezultatami wydarzeń sportowych. Zjawisko to występuje głównie ze strony zorganizowanych grup przestępczych, próbujących ustawić wynik meczu. Widoczne są jednak również przypadki pojedynczych zawodników, którzy działają na mniejszą skalę. Przestępcy działający indywidualnie, najprostszą drogę do wpłynięcia na rezultat zdarzenia sportowego doświadczają w przypadku sportów indywidualnych. Najbardziej dotknięty przez negatywny wpływ manipulacji w dzisiejszych czasach wydaję się tenis. Jeżeli chodzi o tę dyscyplinę sportu, istnieje wiele sposobów, w jaki sposób można by ustawić wynik zdarzenia. Zarówno będąc faworytem, poprzez odpuszczenie meczu w odpowiednim momencie, jak i będąc zawodnikiem nisko klasyfikowanym. Przykładowo, zawodnik zaangażowany w manipulacje, stawiany w roli zdecydowanego faworyta, oprócz możliwości całkowitego odpuszczenia spotkania, może postarać się wygrać pierwszą część (zakładając, że zawodnicy rozgrywają spotkanie do 2 wygranych setów). W takiej sytuacji, podmioty przyjmujące zakłady oferują niezwykle atrakcyjne zwroty, jeśli ich klient zdecyduje się typować zwycięstwo przeciwnika (gdyż prawdopodobieństwo takiego zdarzenia, z punktu widzenia czysto sportowego jest niezwykle niskie). W bardzo łatwy sposób rezultat meczu tenisa ziemnego może zostać wypaczony również poprzez brak zaangażowania zawodnika gorszego. Nie stawiając oporu, może on doprowadzić do wysokiej przegranej, za co także oferowane są wysokie zwroty u firm bukmacherskich.

1.1 Manipulacje w piłce nożnej

Problem manipulacji dotyczy jednak nie tylko rozgrywek określanych jako indywidualne. Każdego dnia na świecie odbywa się również setki spotkań w dziedzinie sportu, jaką jest piłka nożna, spora część z nich jest śledzona przez sympatyków na żywo, czy też poprzez relacje w mediach. Zainteresowanie, a co za tym idzie opłacalność finansowa dla zawodników biorących udział w wydarzeniach, spada wraz ze spadkiem poziomu rozgrywek, mówiąc przykładowo o 3-4. dywizji w danym kraju, można zaryzykować stwierdzenie, iż pojawiający się tam zawodnicy uprawiają ten sport całkowicie hobbystycznie. Pomimo, iż spotkania te nie przyciągają rzeszy fanów, w ofercie firm bukmacherskich bardzo często pojawia się możliwość zawierania zakładów na tego typu zdarzenia. Jak już wspomniano, jest to zrozumiałe z punktu widzenia bukmachera, który rozszerzając ofertę pozyskuje nowych klientów, aczkolwiek otwiera również ogromne możliwości manipulacji spotkaniami w celu osiągnięcia korzyści majątkowych. Każdego roku liczne organizacje (Sport Radar, Genius Sports Group, Stats Perform itp.) monitorujące globalnie zawierane zakłady, publikują raporty dotyczące spotkań. Zawierają one wydarzenia sportowe, w których wzorce, kwoty zawieranych zakładów wskazują na pewne nieprawidłowości. Raporty te wskazują, iż w przypadku piłki nożnej, rocznie odbywa się około 500 spotkań, w których zarejestrowano podejrzane aktywności, jeżeli chodzi o kwoty przyjmowanych zakładów.

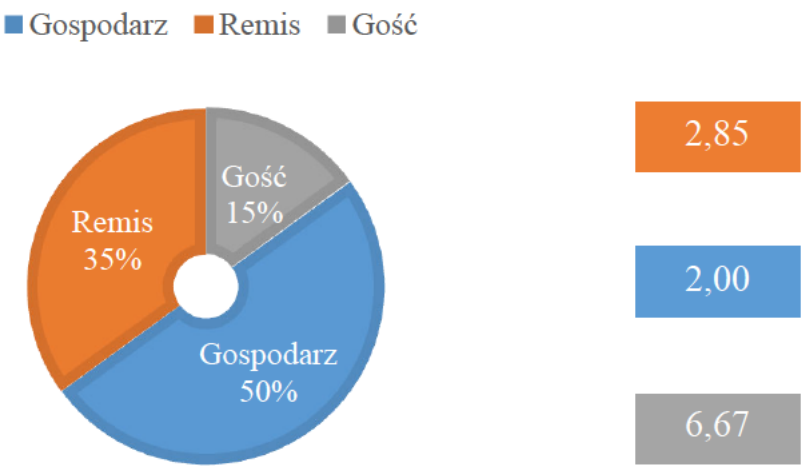
1.2 Prawdopodobieństwo w zakładach bukmacherskich

W realiach zakładów bukmacherskich czasami wytwarzają się różnice w kursach na dane zdarzenie sportowe, wynika to wprost z innej oceny prawdopodobieństwa przez różne podmioty. Teoria dotycząca kursów bukmacherskich, przytoczona w pracy (Cortis, 2015), pozwala lepiej zrozumieć użyte pojęcia oraz dokonane transformacje, widoczne w dalszej części pracy.

Podmiot bukmacherski wystawiając swoją ofertę, wyraża ją poprzez zestawienie odpowiednio przygotowanych kursów na poszczególne zdarzenia, stanowiących informację dla potencjalnych klientów. W Europie najczęściej stosowany jest system dziesiętny, zwany też europejskim, pokazuje on krotność stawki, jaką otrzyma gracz, jeśli jego zakład okaże się wygrany. Dla kontrastu, dla klientów zakładów bukmacherskich

w Wielkiej Brytanii normą jest używanie kursów wprost odzwierciadlających pojęcie szansy na zajście danego zdarzenia. W pracy skupiono się na systemie europejskim.

Od strony bukmachera, kurs to nic innego, jak ocena prawdopodobieństwa danego zdarzenia, z uwzględnieniem swojej marży. Przy jej pominięciu, kurs dziesiętny jest dokładnie odwrotnością oceny prawdopodobieństwa zdarzenia (Rysunek 1).



Rysunek 1. Zamiana prawdopodobieństwa (wyrażonego procentowo) na kurs dziesiętny.
Źródło: Opracowanie własne

W praktyce kursy oferowane przez konkurencyjne podmioty, różnią się o maksymalnie kilka procent w danym momencie, a gdy powstają większe różnice, zaczyna być to interesujące (np. w kontekście arbitrażu), bukmacherzy notując u siebie niezgodność kursów z konkurencją, po upływie chwili wstrzymują przyjmowanie zakładów na dany rynek i analizują poprawność swojej oferty. Analitycy mają na to czas, jeśli mowa o zakładach przedmeczowych, sytuacja ma się zupełnie inaczej, gdy mowa o bardzo popularnych zakładach na żywo.

Spadek kursu w zakładach na żywo, poza oczywistymi, typowo sportowymi powodami wynikającymi z przebiegu meczu (gol, kartka, upływający czas itp.) może być również spowodowany faktem przyjmowaniem na dany rynek podejrzanych zakładów o wolumenie przekraczającym wyznaczone normy – poważniejsze podmioty bukmacherskie korzystają z odpowiednich mechanizmów obniżających kursy na obłożony rynek (rośnie prawdopodobieństwo danego zdarzenia), tym samym rośnie

kurs zdarzenia przeciwnego (spada jego prawdopodobieństwo), co m.in. tworzy okazje arbitrażowe pomiędzy bukmacherami, dzięki czemu wychwytywane jest to przez odpowiednie oprogramowanie.

Celem pracy jest próba stworzenia modelu, który ma za zadanie ocenić prawdopodobieństwo nadużycia w podejrzanym meczu piłki nożnej, zgodnie z kierunkiem zasugerowanym przez wahania kursów bukmacherskich, na podstawie czynników pozasportowych. Zaplanowano użycie technik analizy dyskryminacyjnej oraz regresji logistycznej. Dzięki zastosowaniu wspomnianych metod, oszacowanym prawdopodobieństwom klasyfikacji obserwacji do odpowiednich grup. Wykorzystując prawdopodobieństwa, w dalszej części zaproponowano także system wspomagający decyzje inwestycyjne dotyczące zawierania zakładów na spotkania, w których wytwarzają się wysokie różnice kursowe.

1.3 Charakterystyka oraz wstępna analiza zbioru danych

Analizowany zbiór obserwacji dotyczy spotkań piłki nożnej, w których dochodziło do nagłych, podejrzanych spadków kursów, podczas trwania spotkania. Autor gromadził dane przez okres jednego roku, zapisując najważniejsze, w jego ocenie, czynniki, pozwalające ocenić na ich podstawie, czy doszło do manipulacji w kierunku zgodnym z przewidywaniami. Gromadzenie najistotniejszych w tym wypadku informacji było możliwe dzięki oprogramowaniu betburger, które porównuje ze sobą oferty większości firm bukmacherskich (jak i również giełdy zakładów – betfair) na rynku, w poszukiwaniu różnic kursowych. Wstępny zbiór potencjalnych zmiennych objaśniających można podzielić na dwie grupy, pierwsza z nich dotyczy informacji o ruchach kursów, ich kierunku, intensywności, wprost uzyskanych z użyciem wspomnianego oprogramowania. Zmienne składające się na opisywaną podgrupę:

Drop – jest to informacja, o wysokości spadku kursu bukmacherskiego na zdarzenie, w praktyce, ze względu na ręczny sposób kompletowania obserwacji i specyfikę używanego do tego oprogramowania, wyraża procentowo, jak wysoką okazję arbitrażową utworzył dany spadek kursu, pomiędzy bukmacherami, którzy reagują w takich wypadkach najszybciej (jak już wspomniano – posiadają mechanizmy balansujące wysokość kursów na podstawie łącznego wolumenu przyjętego na dany rynek) , względem podmiotów nieposiadających, na tyle zaawansowanych technologii aby odpowiednio szybko w takich przypadkach reagować.

Używając słowa zdarzenie, autor ma na myśli konkretny zakład w obrębie danego rynku bukmacherskiego w meczu, przykładowo, zdarzeniem (a tym samym kierunkiem spadków) w wybranym spotkaniu mogą być zakłady dotyczące ilości bramek powyżej 2,5 (wartość niecałkowite na rynkach bukmacherskich m.in. dotyczących ilości bramek, powstały w celu uproszczenia koncepcji, przykładowo wartość 2,5 pozwala na dwu-drogową koncepcję rynku, w sytuacji gdy liczby całkowite użyte w takim wypadku tworzą opcje trzy-drogową – powyżej 2 bramek, dokładnie 2 bramki oraz poniżej 2 bramek), a rynkiem w takim wypadku określa się wszystkie zdarzenia zaliczające się do jednej z dwóch kategorii: powyżej/poniżej 2,5 gola.

Alert – zmienna binarna - opiera się głównie na fakcie obecności podejrzanej aktywności na giełdzie zakładów (np. próby gry po mocno zaniżonych kursach) w kierunku zgodnym ze spadkami kursów.

W przypadku, gdy dane spotkanie oferowane było przez giełdę zakładów (57% obserwacji), badacz miał możliwość podglądu konkretnych kwot, za jakie próbowano się zakładać w podejrzanym kierunku w danych meczach. Przyporządkowane jej wartości opierały się więc na subiektywnej ocenie badacza, na podstawie jego doświadczenia i wiedzy, w każdym indywidualnym przypadku.

Druga podgrupa zmiennych dotyczy głównie informacji, statystyk związanych z rodzajem, poziomem rozgrywek, w ramach których toczyły się konkretne spotkania. Wyróżniono trzy takie zmienne:

Cup – zmienna binarna, niosąca informację, czy dany mecz odbywał się w ramach pucharu lub też meczu towarzyskiego.

Event_level – w tym wypadku zmienna w pewnym stopniu informuje o poziomie gorszej z drużyn, między którymi rozgrywał się mecz. Poziom drużyn określony został na podstawie ilorazu dwóch wartości – pozycji danej ligi w zestawianiu wszystkich dywizji, które rozgrywane są w danym kraj, względem ilości wszystkich lig, należących do danej federacji. Ilorazy te obliczane są zarówno dla drużyny gospodarzy, jak i gości, by ostatecznie wybrać większa z nich (im wartość bliższa 1, tym niższy poziom drużyny), w dużej części przypadków oba ilorazy wyniosły tyle samo (mecze ligowe), lecz wybrano taki sposób reprezentacji, głównie po to, aby uwzględnić wpływ faktu pochodzenia drużyn z niskich lig np. w przypadku, gdy rozgrywają one sparing z drużynami z wyższych dywizji, na występowanie

manipulacji. Warto wspomnieć, iż do ustawienia spotkania bardzo często wystarczy pojedynczy zawodnik, czy też sędzia, a w sytuacji, gdy w manipulację zaangażowana jest cała drużyna, przeciwnik może nawet nie zdać sobie sprawy z faktu brania udziału w spotkaniu o charakterze podejrzanym, stąd chęć uwzględnienia poziomu słabszej drużyny, wśród możliwości której leży przykładowo celowe odpuszczenie meczu, pozwalając zdobyć rywalowi dużo bramek.

W Tabeli 1 przedstawiono przykład obliczania wartości zmiennej **Event_level**, na jednej z obserwacji ze zbioru, jej wartość w przypadku meczu pochodzącego z rozgrywek Greece Cup, został oszacowany na 0,4, na co składa się występowanie drużyny Apollon Larissa F.C, grającej na drugim, z pięciu szczebli greckiego systemu rozgrywek, wraz z Xanth F.C, pochodzącej z krajowej ekstraklasy.

Gospodarz	Gość	Liga	Poziom gospodarza	Poziom gościa	Event_level
Larissa	Xanthi	Greece Cup	2/5	1/5	2/5 (0,4)

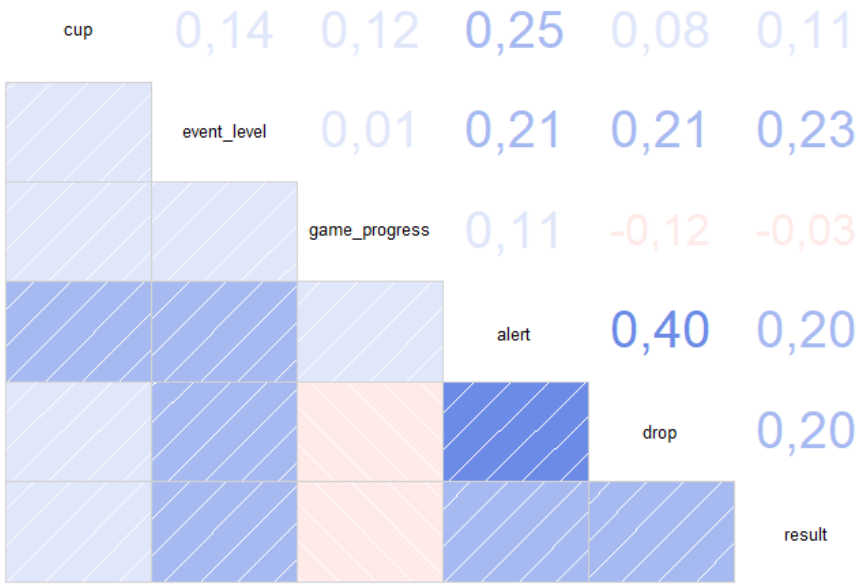
Tabela 1. Sposób obliczania zmiennej event_level.

Źródło: Obliczenia własne

Ostatnią zmienną, która znalazła się w zbiorze jest **game_progress**, określa ją stosunek numeru rundy, z którego pochodził mecz, względem ilości rund zaplanowanych w ramach danych rozgrywek, aczkolwiek wartości tej zmiennej w niektórych wypadkach nie mogły zostać ustalone, ze względu na brak dostępności odpowiednich danych (najniższe ligi w krajach azjatyckich, czy też mecze towarzyskie). W przypadku pucharów, na ile było to możliwe, przyporządkowywano, w podobny sposób, wartość odpowiadającą postępowi rozgrywek.

Problem braków danych w zbiorze dotyczył jedynie zmiennych **game_progress** oraz **event_level**. Pierwszy z wspomnianych predyktorów nie zawierał wystarczającej informacji w przypadku 55 obserwacji. Zmienna **event_level** obarczona była brakiem danych dla 26 spotkań. Brakujące wartości zostały zastąpione średnią arytmetyczną.

W kolejnej części pracy, skupiono się na opisaniu podstawowych statystyk dotyczących badanego zbioru zmiennych.



Rysunek 2. Zestawienie korelacji zmiennych.
Źródło: Opracowanie własne (Rstudio)

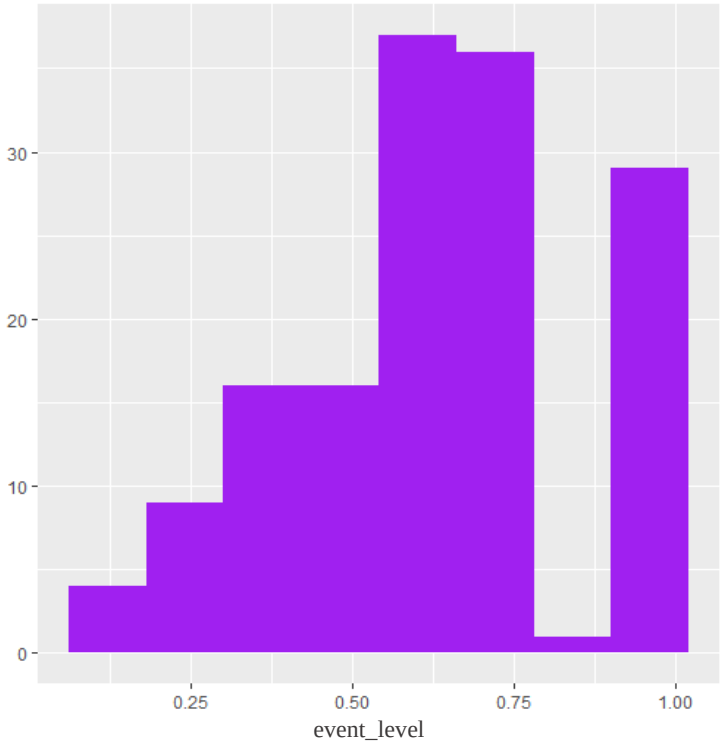
Wybrane zmienne objaśniające wykazują niskie korelacje między sobą (Rysunek 2), wyróżnia się jedynie powiązanie zmiennej **alert** z wartościami **drop**, co wydaje się całkowicie wytłumaczalne, gdyż podejrzanym zagraniom na giełdzie zakładów bardzo często towarzyszą istotne spadki kursów. W ostatniej kolumnie macierzy, odnoszącej się do powiązań zmiennej objaśnianej z poszczególnymi predyktorami, uwagę zwraca wartość -0,03, dotycząca zmiennej **game_progress**, sugerująca jej minimalny (ujemny) wpływ na wynik, co sugerować mogłoby jej pominięcie już na tym etapie. Warto jednak dodać, iż w dalszej części pracy podjęta została próba doboru takich zmiennych, których łączny wpływ ma, jak największe znaczenie w kontekście przynależności obserwacji do odpowiedniej grupy, a więc nie należało na tym etapie, na podstawie macierzy korelacji, żadnej ze zmiennych odrzucać.

zmienna	zmienność	średnia	odch. stand
cup	1,543	0,297	0,459
event_level	0,353	0,649	0,229
game_progress	0,355	0,598	0,212
alert	1,770	0,243	0,430
drop	0,525	0,311	0,164
result	0,706	0,669	0,472

Tabela 2. Podstawowe statystyki zmiennych ze wstępnego zbioru.

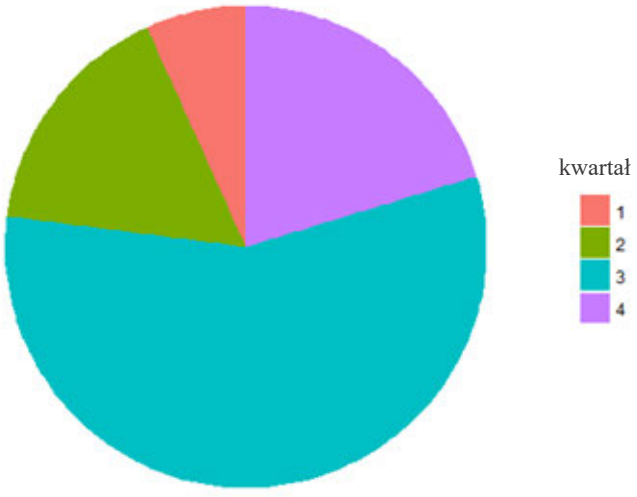
Źródło: Obliczenia własne

Analiza Tabeli 2, pozwala stwierdzić, iż w badanym zbiorze obserwacji, zmienność kształtuje się na satysfakcjonującym, w kontekście budowy dalszego modelu, poziomie, w każdym z przypadków. W analizowanym zbiorze około 30% obserwacji dotyczyło spotkań rozgrywanych w ramach pucharów lub też spotkań towarzyskich, z kolei 24% meczów w opinii badacza charakteryzowało się nieregularnymi ruchami na giełdzie zakładów. Średni spadek kursu pośród badanych obserwacji oscylował w okolicach 32%. Proporcja meczów w badanym zbiorze obserwacji, w których doszło do nadużycia zgodnie z kierunkiem zmiany kursów, względem reszty, wynosi około 2:1. Ciekawym faktem wydaje się również informacja, iż podejrzane spotkania toczyły się średnio na etapie ukończenia rozgrywek na poziomie 60%, a wartości tej zmiennej odchylały się przeciętnie od średniej o 0,22j. Zdecydowana większość analizowanych meczów odbywała się w końcowych etapach rozgrywek (Rysunek 4). Wśród zebranych obserwacji zauważono również wyraźny trend do występowania w tego typu meczach drużyn z niskich lig. W około trzydziestu przypadkach (20%) jedna z drużyn pochodziła z najniższej ligi w swoim kraju, co obrazuje ostatni słupek z wykresu (Rysunek 3).



Rysunek 3. Histogram dla zmiennej event_level.

Źródło: Opracowanie własne (Rstudio)



Rysunek 4. Rozkład podejrzanych meczów w poszczególnych częściach sezonu.

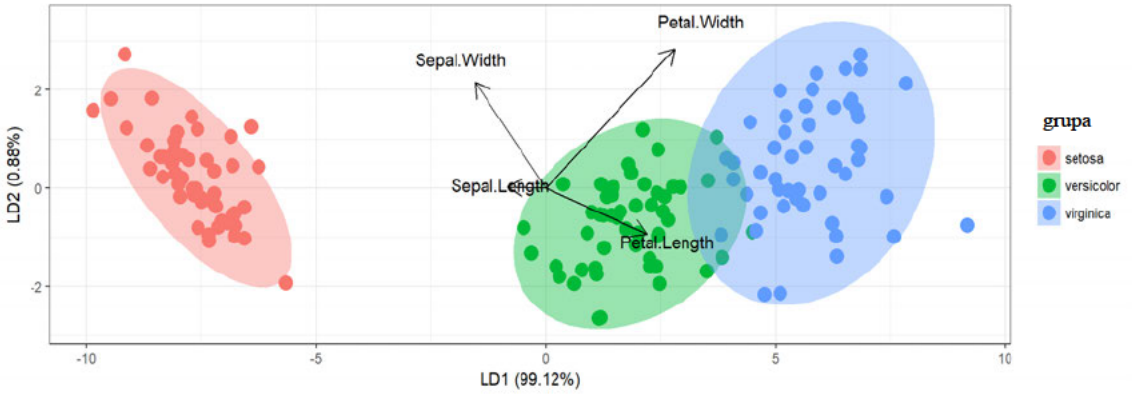
Źródło: Opracowanie własne (Rstudio)

2. Techniki badawcze

Badacz wykorzystał w pracy dwie techniki związane z problemem klasyfikacji obserwacji do odpowiednich podgrup. Skorzystano z analizy dyskryminacyjnej oraz regresji logistycznej. Ze względu na binarny charakter zmiennej objaśnianej, skupiono się na użyciu obu technik do klasyfikacji w kontekście jedynie dwóch grup. Rozdział ten zawiera teoretyczny wstęp dotyczący wykorzystanych metod.

2.1 Analiza dyskryminacyjna

Analiza dyskryminacyjna w statystyce określana jest jako metoda realizująca dwa podstawowe zadania: badanie różnic pomiędzy obserwacjami należącymi do poszczególnych grup (dwóch lub więcej), jak i również konstruowanie możliwie najlepszej reguły przydziału obserwacji o nieznanej klasie, do już istniejącej (Koronacki i Ćwik, 2005). Problematykę tą zainicjował znany matematyk - Ronald Fischer, autor również innej, powszechnie stosowanej analizy statystycznej, dotyczącej wariancji. W stworzonym przez Fischera algorytmie liniowej analizy dyskryminacyjnej (LDA) koncepcja oparta jest na utworzeniu reguły dyskryminacyjnej na podstawie próby uczącej, przy wykorzystaniu funkcji liniowej. Reguła ta ma za zadanie przyporządkować nowe obserwacje do jednej z grup (z góry określonych). Warto zaznaczyć, iż w podejściu Fischera dopuszcza się podział obserwacji na jedynie dwie grupy.



Rysunek 5. Przykładowy podział obiektów (IRIS), ze względu na wartości funkcji dyskryminacyjnych LD1 i LD2. Źródło: Opracowanie własne (Rstudio)

Wspomniane kryterium przyporządkowania obserwacji do grup, w postaci liniowej funkcji dyskryminacyjnej, wyglądać może następująco (Radkiewicz, 2010):

$$Y_m = a_0 + a_1 X_{1m} + a_2 X_{2m} + .. + a_{km} X_{km}$$

Y_{mj} – wartość funkcji dyskryminacyjnej dla obserwacji m

a_i – współczynniki funkcji dyskryminacyjnej

X_{im} – wartość i-tej zmiennej dyskryminacyjnej dla obserwacji m

k – liczba zmiennych

a_0 – stała

Współczynniki dyskryminacyjne określają wpływ poszczególnych zmiennych na funkcję, istotą ich estymacji jest otrzymanie takiej funkcji, dzięki której obserwacje pomiędzy określonymi grupami będą maksymalnie wyizolowane. Ilość funkcji dyskryminacyjnych zależy m.in. od liczności zbioru grup, do których należą obserwacje.

Główne założenia omawianej metody są następujące:

- wielowymiarowy rozkład normalnych danych,
- równość kowariancji w grupach,
- brak zjawiska współliniowości wśród zmiennych objaśniających,
- odpowiednia ilość obserwacji (większa od ilości zmiennych o co najmniej dwa),
- liczność każdej z grup na poziomie dwa lub więcej,
- obserwacje należące do grup wzajemnie się wykluczających.

Odnosząc się do dwóch początkowych punktów - automatycznie uniemożliwiają one stosowanie w modelach zmiennych objaśniających o charakterze jakościowym, aczkolwiek badacze podkreślają (Koronacki i Ćwik, 2005), iż analiza dyskryminacyjna potrafi dawać sensowne rezultaty w poszczególnych przypadkach także dla danych niespełniających tych założeń.

Jedną z podstawowych statystyk służących ocenie modelów stworzonych z użyciem LDA może być λ-Wilksa, która w przypadku istnienia jedynie dwóch grup obserwacji wyraża stosunek wewnątrzgrupowej i całkowitej sumy kwadratów funkcji dyskryminacyjnej. Statystyce tej towarzyszy test hipotezy zerowej o braku różnic międzygrupowych w próbce. Weryfikacja trafności modelu odbywa się także poprzez określenie odsetka obserwacji poprawnie zaklasyfikowanych, również z wyszczególnieniem tych wartości wśród grup.

2.2 Regresja logistyczna

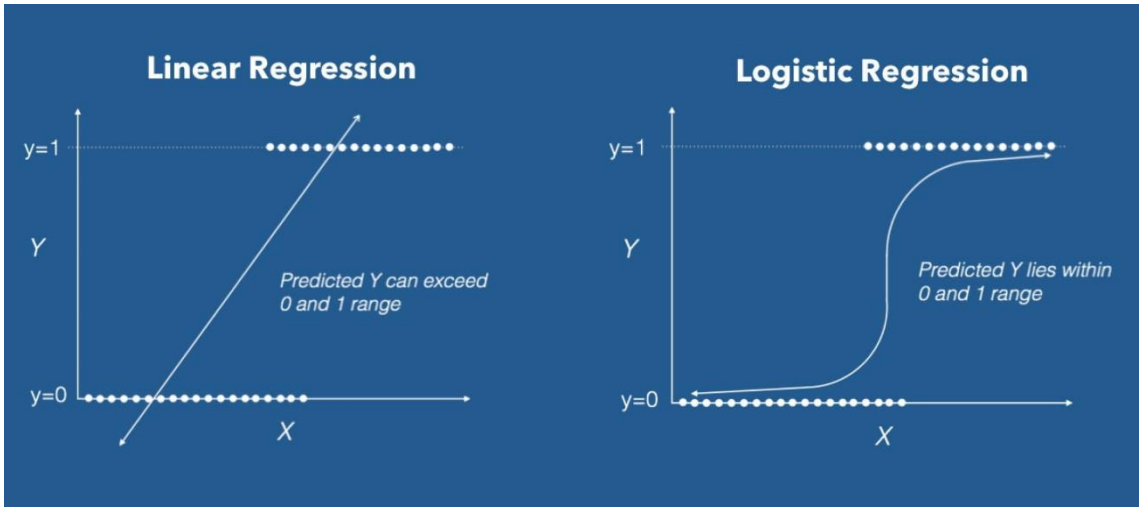
Model regresji logistycznej to jeden z najczęściej używanych algorytmów w kategorii problemów klasyfikacyjnych. W podstawowej wersji, stosowany jest, gdy zmienna objaśniana reprezentowana jest przez dokładnie dwie wartości. Dodatkowym faktem, różnicującym ten rodzaj regresji od jej liniowej odmiany, jest możliwość otrzymania na podstawie zestawu predyktorów, prawdopodobieństwa (p) pewnego zdarzenia, zamiast przewidywania konkretnej wartości.

Według autorów książki (Bruce i Bruce, 2007), mówiąc o regresji logistycznej, w podejściu naiwnym, można by ją postrzegać w postaci funkcji liniowej:

$$p = \beta_0 + \beta_1 x_1 + .. + \beta_n x_n$$

Podejście takie, nie daje jednak gwarancji, iż wartość p znajdzie się w przedziale <0;1> (Rysunek 6). By to osiągnąć, modelowanie p odbywa się poprzez zastosowanie funkcji logistycznej. Przekształcenie to, pozwala wyrazić prawdopodobieństwo w następujący sposób

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + .. + \beta_n x_n)}}.$$



Rysunek 6. Regresja liniowa i logistyczna.
Źródło: (Prabhakaran, 2017)

Prawdopodobieństwo w przypadku tego rodzaju regresji, rozważa się w głównie postaci ilorazu szans (ang. odds ratio). Z praktycznego punktu widzenia, jest to iloraz, który

wyraża prawdopodobieństwo na to, iż dane wydarzenia się wydarzy w stosunku do prawdopodobieństwa, że dane zdarzenie nie będzie miało miejsca (Danieluk, 2010).

Po zastosowaniu odpowiednich przekształceń oraz nałożeniu logarytmu naturalnego na obie strony równania, zależność logarytmu wspomnianego ilorazu szans od zestawu predyktorów, wyraża się jako:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + .. + \beta_n x_n$$

Dzięki liniowej zależności logarytmu szans od zmiennych x_i, interpretacja współczynników przy nich stojących sprowadza się do mówienia o wpływie zmiany o jednostkę czynnika, wyrażonego przez x_i na zmianę wartości logarytmu. Jeśli chodzi o rozpatrywanie terminu ilorazu szans, wyrażenia e^{β_i} zwiększają prawdopodobieństwo wystąpienia rozpatrywanego zjawiska, w momencie, gdy osiągają wartości większe niż 1, zachowują neutralny wpływ przy poziomie równym liczbie jeden, a w pozostałych przypadkach mówi się o ograniczającym wpływie na badane zjawisko.

O wpływie jednostkowej zmiany wartości zmiennej wspomina (Hosmer i Lemeshow, 2000), aczkolwiek w celu obliczenia wpływu zmiany danego czynnika o dowolną wartość a, na iloraz szans, należy zauważyć pewne przekształcenie:

$$e^{\beta_0 + \beta_i(x_i+a) + .. + \beta_n x_n} = e^{\beta_0 + \beta_i x_i + .. + \beta_n x_n} \cdot e^{a \cdot \beta_i} = \frac{p}{1-p} \cdot e^{a \cdot \beta_i}$$

Przytoczony wzór pozwala wnioskować, iż wzrost wartości predyktora o a, zwiększa $e^{a \cdot \beta_i}$ razy wartość ilorazu szans.

W odróżnieniu od liniowej analizy dyskryminacyjnej, model regresji logistycznej nie jest szczególnie ograniczany przez założenia. Zmienne objaśniające nie muszą wykazywać się rozkładem normalnym, dopuszcza się również heteroskedastyczność. Przyjmuje się jednak pewne warunki:

- brak współliniowości wśród zmiennych objaśniających,
- liniowa zależność logitu prawdopodobieństwa od predyktorów.

W kategoriach oceny jakości dopasowania modeli logitowych, stosowane są miary bazujące na odchyleniu modelu (ang. model deviance). Jedną z nich jest kryterium Akaike, które przyjmuje tym mniejszą wartość, im lepsze dopasowanie. Kryterium AIC określa jako:

$$AIC = dev_{\omega} + 2\,(p+1),$$

gdzie dev_{ω} to dewiancja badanego modelu, a p + 1 to liczba parametrów modelu. Obciążenie o podwojoną liczbę parametrów jest stosowane ze względu na oczywisty wpływ zmniejszania się odchylenie, w przypadku dodawania nowych zmiennych do modelu.

Wykorzystuje się również test Walda, analizujący, w najprostszym wypadku (testowanie pojedynczego parametru), iloraz oszacowania parametru β_i oraz błędu standardowego jego oszacowania. Hipoteza zerowa tego testu stanowi o nieistotności badanego parametru. (Koronacki i Ćwik, 2005) podkreślają tu jednak, iż nieodrzucenie hipotezy zerowej w tym wypadku, nie oznacza braku wpływu zmiennej X_i, przy której stoi parametr β_i, na zmienna objaśnianą Y, a jedynie sugeruje, iż wpływ ten został wyjaśniony przez pozostałe predyktory.

3. Wyniki przeprowadzonych badań

Trzeci rozdział pracy dotyczy prezentacji otrzymanych wyników przy użyciu zastosowanych metod. Podjęto działania w celu wyłonienia modelu optymalnego w kontekście, jak największej trafności, jeśli chodzi o predykcję. Następnie zestawiono wyniki otrzymane przy użyciu obu technik. Zaproponowano także sposób wykorzystania modelu we wspomaganiu decyzji inwestycyjnych.

3.1 Analiza dyskryminacyjna

Ze względu na założenia analizy dyskryminacyjnej, w początkowej fazie wyodrębniono ze zbioru zmiennych objaśniających te o charakterze ilościowym – **drop, event_level, game_progress**.

	Lambda Wilksa: 0,921; p-value < 1%	
	Lambda Wilksa	p-value
event_level	0,959	0,015
drop	0,945	0,051

Tabela 3. Wynik działania analizy krokowej postępującej dla zmiennych: drop, event_level, game_progress.
Źródło: Obliczenia własne

Tabela 3 przedstawia m.in. poziomy istotności poszczególnych zmiennych. Z modelu podczas analizy krokowej odrzucono zmienną **game_progress**, której wpływ na różnicowanie grup, w odróżnieniu od zmiennych pozostałych w modelu, był znikomy. Statystyka λ-Wilksa dla całego modelu ukształtowała się na poziomie około 0,92, co sygnalizowało bliskość średnich między grupami dla rozpatrywanego zestawu zmiennych. Model dodatkowo został sprawdzony pod kątem reklasyfikacji obserwacji, uzyskano skuteczność na poziomie 67%. Zasugerowało to badaczowi konieczność dalszej pracy nad modelem.

Zdecydowano sprawdzić model zawierający wszystkie zmienne z początkowego zbioru danych, oczekując otrzymania modelu o większej skuteczności. Po zastosowaniu analizy krokowej na zaproponowanym zbiorze, otrzymano model o zbliżonej trafności 65%, składały się na niego zmienne: **event_level, alert, drop**. Średnie wartości poszczególnych zmiennych wśród grup w tym wypadku również nie różnicowały ich wystarczająco.

	Lambda Wilksa: 0,911; p-value < 1%	
	Lambda Wilksa	p-value
event_level	0,943	0,025
drop	0,922	0,178
alert	0,921	0,206

Tabela 4 Wynik działania analizy krokowej dla wszystkich zmiennych z początkowego zbiory danych.

Źródło: Obliczenia własne

W kolejnym kroku postanowiono zbadać dokładniejszy wpływ faktu rozgrywania spotkań w obrębie konkretnych lig na występowanie nadużyć. W tym celu dla każdej ligi, z której występowały obserwacje stworzono zmienną binarną informującą, czy mecz dotyczył tej ligi. Odrzucono zmienne dotyczące tych lig, wśród których występowały mniej niż trzy obserwacje. Na model złożyły się nowe zmienne w połączeniu ze zmiennymi z początkowego zbioru danych, przy zaznaczeniu, iż pominięto te, w których średnie wartości nie różniły się pomiędzy grupami. Działania te pozwoliły uzyskać badaczowi wzrost odsetka poprawnie przydzielonych obserwacji, podczas reklasyfikacji. Ostatecznie otrzymano model, w którego skład weszły zmienne widoczne w Tabeli 5.

zmienna	współczynnik LD1	stand. współczynnik
cup	-0,282	-0,129
event_level	2,334	0,521
alert	0,524	0,223
drop	1,234	0,198
league.Armenia_1st_Divisiom	-0,134	-0,027
league.Azerbeijan_1st_Division	-2,433	-0,435
league.Brazil_u20_Cup	-0,792	-0,129
league.Bulgarian_u19	-2,049	-0,289
league.Club_Friendlies	0,486	0,167
league.Czech_3league	-0,850	-0,153
league.Greece_Football_League	0,434	0,104
league.India_Goa_Pro_League	-0,085	-0,022
league.Indonesia_Lower_Leagues	-1,901	-0,377
league.Russia_2nd_League	-2,393	-0,470
league.Vietnam_Second_Division	0,392	0,094
league.Vietnam_u19_league	0,247	0,040
constant	1,709	

Tabela 5. Zmienne tworzące model wraz z współczynnikami funkcji dyskryminacyjnej.

Źródło: Obliczenia własne

Na finalny model, otrzymany poprzez wykorzystanie liniowej analizy dyskryminacyjnej złożyło się szesnaście predyktorów, wagi poszczególnych w równaniu funkcji dyskryminacyjnej, określone zostały poprzez oszacowane współczynniki LD1 (Tabela 5). W ten sposób otrzymano informację o wpływie zmiany wartości predyktora o jednostkę, na wartość funkcji dyskryminacyjnej. Po przeanalizowaniu standaryzowanych współczynników, określono siłę oddziaływania poszczególnych predyktorów. Okazało się, iż zmienne **event_level**, wraz z trzema predyktorami dotyczącymi lig, w ramach których rozgrywał się mecz (**Russia_2nd_League**, **Azerbeijan_1st_Division** oraz **Indonesia_Lower_Leagues**) miały największy wpływ na wartość funkcji dyskryminacyjnej danej obserwacji.

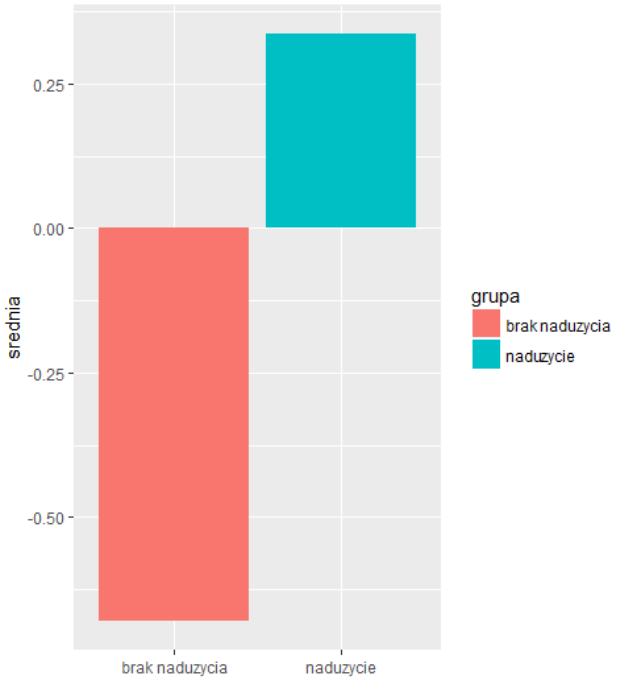
liga	cup	event_level	alert	drop	result	prog. wynik	LD1	P_Gr_0	P_Gr_1
Greece F-L	0	0,600	0	0,150	1	1	0,310	23%	77%
Isreal Liga	0	0,600	0	0,300	1	1	0,061	28%	72%
Russia 2	0	0,750	1	0,350	1	0	-1,395	63%	37%
Russia 2	0	0,750	1	0,250	0	0	-1,519	66%	34%
FA Cup	1	0,375	0	0,080	0	0	-1,017	54%	46%

Tabela 6. Zestawienie kilku obserwacji wraz z prognozami, wartościami LD oraz prawdopodobieństwami przynależności do grup.

Źródło: Obliczenia własne

Przedstawione wyniki pierwszych pięciu obserwacji ze zbioru danych (Tabela 6), w połączeniu z wartościami z Rysunek 7 pozwalają określić obserwację pierwszą, jako typowego przedstawiciela grupy nr 1, związanej z meczami, w których prawdopodobnie doszło do nadużyć. Wynik dyskryminacyjny bowiem kształtuje się dla tego przypadku na poziomie 0,31, czyli bardzo blisko średniej grupowej. Model określa trafność tej klasyfikacji na około 77%. Jeśli chodzi o spotkanie w ramach FA Cup, przedstawione na spodzie tabeli, prognoza okazała się trafna, aczkolwiek z dużą szansą na nieprawidłową klasyfikację. Warto zauważyć, iż wynikać to może, z faktu, że zmienna drop, wyrażająca procentowy spadek kursu w tym wypadku ma bardzo niską wartość, co w połączeniu z dość wysokim współczynnikiem LD1 dla tego predyktora (Tabela 5), zadecydowało o takim wyniku. Dodatkowo, sytuacja taka dobrze świadczy o modelu, ponieważ poza mierzalnymi wskaźnikami służącymi do jego oceny, ważne jest

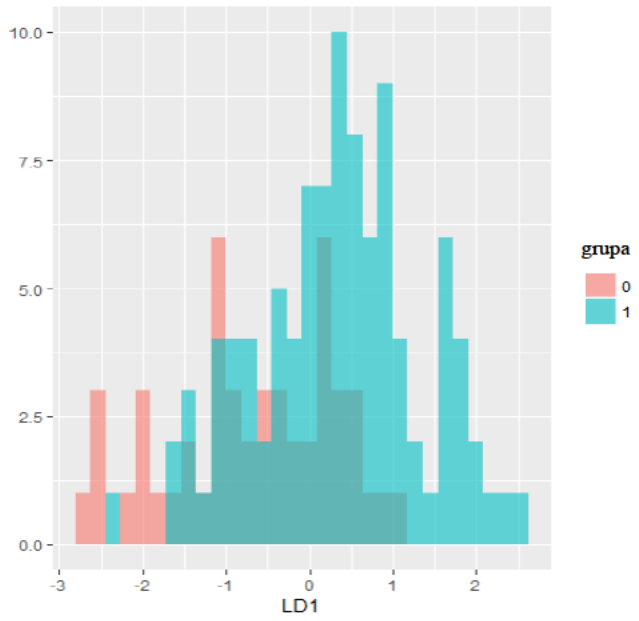
też jak zachowuje się w indywidualnych przypadkach, a w spotkaniach o takiej charakterystyce najczęściej do manipulacji nie dochodzi[1].



Rysunek 7. Średnie wyniki dyskryminacyjne w poszczególnych grupach.
Źródło: Opracowanie własne (Rstudio)

Postanowiono również zobrazować model w postaci histogramów dotyczących obu grup (Rysunek 8). Ogólnie rzecz biorąc, mecze zmanipulowane miały relatywnie wysokie wyniki dyskryminacyjne, a te z drugiej grupy – niskie. Niestety klasy nie zostały wyizolowane w sposób wystarczająco dobry. Widoczne było spore nałożenie się wartości funkcji, w centralnej części grafiki.

[1] Teza oparta na wiedzy i doświadczeniu badacza.



Rysunek 8. Histogramy wyników dyskryminacyjnych w grupach.
Źródło: Opracowanie własne (Rstudio)

Jeśli chodzi o reklasyfikację, omawiany wariant zmiennych objaśniających pozwolił uzyskać poprawną klasyfikację w 72% przypadków. Na wspomnianą trafność, składały się wartości, takie jak: około 45% procentowa poprawność przydzielenia zmiennej **result** wśród grupy „0" oraz 85% w przypadku przeciwnym. Model wykazał tendencję do błędnego przyporządkowywania obserwacji z grupy „0" do grupy „1", jeżeli chodzi o przypadek odwrotny – sytuacja miała się znacznie lepiej (Tabela 7).

		Predykacja	
	Wynik	0	1
Stan faktyczny	0	22	27
	1	14	85

Tabela 7. Reklasyfikacja obserwacji przy użyciu modelu uwzgledniającego ligi, z których pochodzą obserwacje. Trafność - 72%.
Źródło: Opracowanie własne (Rstudio)

Obliczona została również korelacja kanoniczna, określająca związek pomiędzy wartościami funkcji dyskryminacyjnej, a zmienną **result**, w tym wypadku badacz uzyskał wartość 0,44, czyli udało się znaleźć funkcję dość powiązaną z wartościami objaśnianymi. Przechodząc do statystyki użytej do badania poprzednich propozycji modeli, wartość λ - Wilksa oscylowała w okolicach 0,8.

Podjęto się również próby analizy obserwacji, dla których wartości zaproponowanych przez LDA prawdopodobieństw oscylowały blisko progu rozdzielającego grupę „0" oraz

„1” (Tabela 8). Główny wniosek z niej płynący stanowi, iż przypadki graniczne pochodzą głównie z lig, dla których zebrano małą liczbę podejrzanych spotkań, uwypukla to ogólny problem dotyczący badań w tym temacie, gdyż takie mecze są najczęściej rozproszone wśród bardzo wielu różnych rozgrywek. Ważną kwestią w tym wypadku jest także wartość zmiennej **alert.** W ocenie badacza ma ona bardzo duży wpływ, jeśli chodzi o konieczność klasyfikacji meczy do grupy, w której wystąpiło nadużycie. W każdym z przypadków wątpliwych, w kolumnie dotyczącej tej zmiennej, widnieje „0”. Taka sytuacja wpływa na brak wyraźnego przyrostu wartości funkcji dyskryminacyjnej, a tym samym implikuje przynależność do obserwacji granicznych.

liga	cup	event_level	alert	drop	result
Greece Cup	1	0,40	0	0,20	1
Bosnia-Herzegovina Premier League	0	0,17	0	0,35	1
Brazil U20 cup	1	0,65	0	0,40	1
Brazil U20 cup	1	0,65	0	0,25	0
Uganda Cup	1	0,34	0	0,25	0
Indonesia Lower Leagues and Cups	0	1	0	0,35	1
Nikuragua Primiera Division	0	0,33	0	0,04	0
Nikuragua Primiera Division	0	0,33	0	0,12	1

Tabela 8. Obserwacje, których wartości prawdopodobieństw aposteriori były w przedziale (0,47;0,53).
Źródło: Opracowanie własne (Rstudio)

Podsumowując ten etap badań, model uzyskany przy użyciu analizy dyskryminacyjnej pozwolił uzyskać dość wysoki odsetek ponownej klasyfikacji zbioru uczącego, napotykał natomiast trudności przy reklasyfikacji obserwacji oryginalnie należących do grupy „0”. Specyfika zebranych obserwacji, w połączeniu z pewnymi ograniczeniami płynącymi z faktu wyboru wspomnianej metody nie pozwalała uzyskać wystarczająco dobrych rezultatów. Postanowiono podjąć próbę estymacji modelu przy pomocy innych technik, które również dotyczą binarnej zmiennej objaśnianej.

3.2 Regresja logistyczna

Początkowo, analizie poddano model zawierający wszystkie zmienne ze wstępnego zbioru (Tabela 9). Biorąc pod uwagę te predyktory, uzyskano sytuację, w której jedynie zmienną **event_level** można było uznać za istotną, na podstawie testu Walda (zakładając wartość progową na 0,05).

Model 1	AIC: 184,29			
	Estimate Std	Std. Error	z value	Pr(>|z|)
constant	-1,148	0,860	-1,336	0,182
cup	0,381	0,435	0,877	0,381
event_level	1,866	0,861	2,167	0,030
game_progress	-0,331	0,845	-0,391	0,695
alert	0,665	0,536	1,241	0,215
drop	2,182	1,488	1,466	0,143

Tabela 9. Podsumowanie Modelu 1.

Źródło: Obliczenia własne

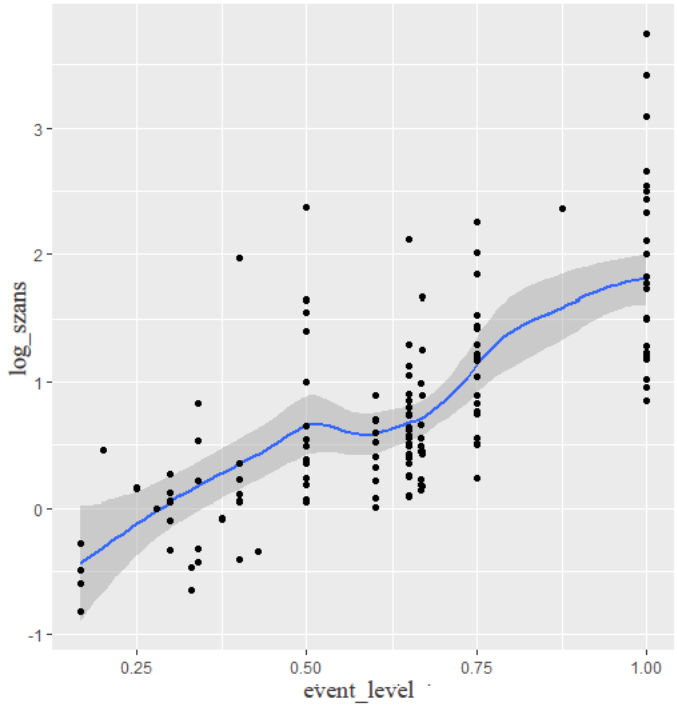
W początkowej fazie rozważań dopasowywania najlepszego modelu, należy również wspomnieć o wartości kryterium AIC, ukształtowała się na poziomie 184,29. Podjęto działania w celu uzyskania lepszego modelu. Następny krok dotyczył utworzenia modeli, uwzględniając każdą zmienną pojedynczo (Tabela 10).

model	AIC	p-value zmiennej	współczynnik
result ~ cup	190,03	0,175	0,547
result ~ event_level	183,67	0,006	2,281
result ~ game_progress	191,83	0,737	-0,279
result ~ alert	185,61	0,020	1,137
result ~ drop	185,25	0,017	3,124

Tabela 10. Modele otrzymane z estymacji przy użyciu każdej zmiennej pojedynczo.

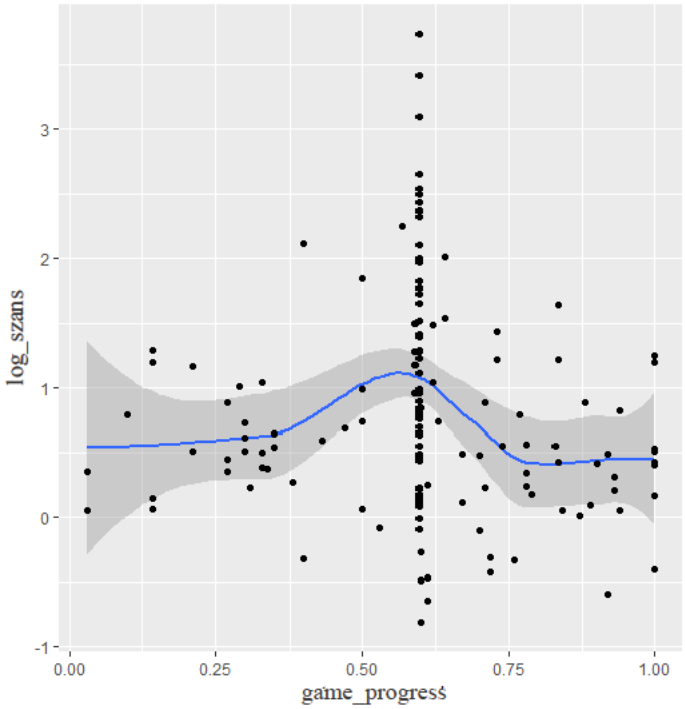
Źródło: Obliczenia własne

Zdecydowanie najgorszą zmienną, w kontekście tworzenia optymalnego modelu, okazał się predyktor informujący o postępie rozgrywek, potwierdzają to: wysoka wartość p-value w Modelu 1 (0,6954), jak i również w przypadku testowania pojedynczego wpływu tej zmiennej (0,737), w połączeniu z brakiem spełnienia założenia o liniowej zależności od wartości tej zmiennej od logarytmu szans (Rysunek 10, Rysunek 11). Przedstawione czynniki zdecydowanie wykluczają rozważaną zmienną z dalszych rozważań. Warto dodać, iż duży wpływ na taki rozwój sytuacji mógł mieć fakt, iż 1/3 obserwacji w obrębie tej zmiennej została zastąpiona średnia, ze względu na brak możliwości uzyskania realnej informacji. Wykluczenie jej z dalszych badań na tym etapie w takim razie, w żadnym wypadku nie dyskwalifikuje wpływu okoliczności odbywania spotkań w końcowych fazach, na podniesione prawdopodobieństwo nadużycia, co znajduje potwierdzenie m.in. w raporcie (2019).



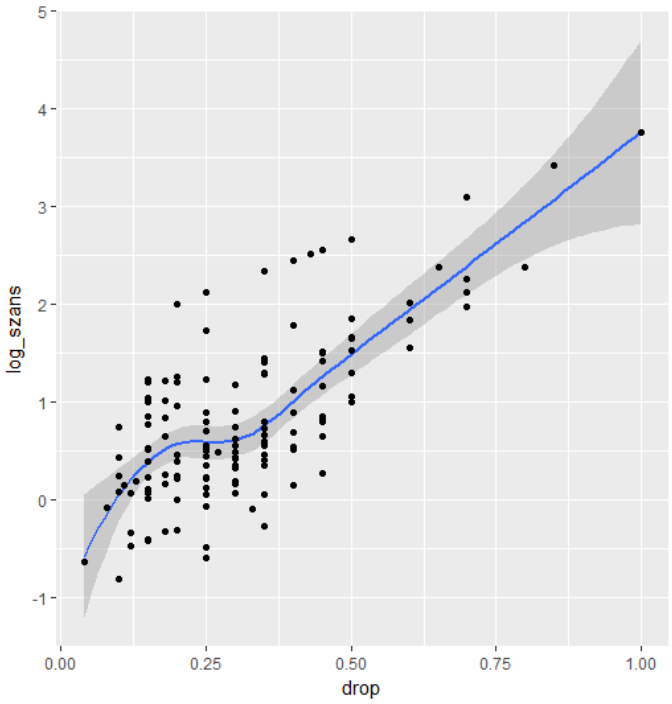
Rysunek 9. Wykres wartości logarytmu szans w zależności od zmiennej event_level.

Źródło: Opracowanie własne (Rstudio)



Rysunek 10. Wykres wartości logarytmu szans w zależności od zmiennej game_progress.

Źródło: Opracowanie własne (Rstudio)



Rysunek 11. Wykres wartości logarytmu szans w zależności od zmiennej drop.
Źródło: Opracowanie własne (Rstudio)

Wykresy (Rysunek 9, Rysunek 11), w przypadku pozostałych dwóch zmiennych ilościowych, nie sugerowały trudności ze spełnieniem założenia o liniowej zależności. Zadecydowano o pozostawieniu ich w dalszych rozważaniach. Przechodząc do zmiennych jakościowych, wartość p-value dla zmiennej **drop** osiągnęła wartość 0,1425, co wykracza poza ogólnie przyjętą normę. Niektóre źródła (Zhongheng, 2016) proponują przesunięcie tej granicy do poziomu 0,25. Uwzględniając ten fakt, zadecydowano uwzględnić omawianą zmienną w dalszym eksperymencie. Kierując się tym kryterium, zdecydowano również o pozostawieniu predyktora **alert** oraz pozbyciu się zmiennej binarnej **cup.**

W wyniku dotychczasowych rozważań otrzymano model składający się z trzech elementów oraz stałej (Tabela 11).

Model 2	AIC: 181,18			
	Estimate Std	Std. Error	z value	Pr(>|z|)
constant	-1,253	0,643	-1,947	0,052
event_level	1,876	0,853	2,201	0,028
alert	0,715	0,521	1,373	0,169
drop	2,147	1,440	1,491	0,136

Tabela 11. Podsumowanie Modelu 2.

Źródło: Obliczenia własne

W porównaniu z Modelem 1, nie uzyskano dużej poprawy, wartość kryterium AIC nieznacznie spadła, po zbadaniu zdolności do odtworzenia próby uczącej, otrzymano odsetek poprawnej reklasyfikacji na poziomie około 67%. W tym wypadku, wyniki estymacji sugerują dodatni wpływ na fakt wystąpienia manipulacji wszystkich trzech czynników: związanego z wysokością spadku kursu, występowania nieregularnych zakładów na giełdzie (**alert**), jak i również zmiennej **event_level**.

Zaproponowano również estymację modelu ze zmiennych, które pozwoliły uzyskać najlepszy rezultat przy użyciu analizy dyskryminacyjnej. Było to poszerzenie podstawowego zestawu zmiennych, o binarne predyktory dotyczące przynależności spotkania do jednej z lig. Zastosowano metodę step (w tym wypadku metoda krokowa wsteczna), dostępną w środowisku R, do wyboru najlepszego podzbioru (Tabela 12).

Model 3	AIC: 173,80			
	Estimate Std	Std. Error	z value	Pr(>|z|)
constant	-2,788	0,934	-2,985	0,003
cup	1,443	0,805	-1,794	0,073
event_level	4,952	1,410	3,511	0,001
drop	2,428	1,605	1,513	0,130
leagues.AFC_CUP	1,726	1,063	1,624	0,104
leagues.Azerbaijan_First	-1,959	1,167	-1,678	0,093
leagues.Bosnia_Premier_League	2,099	1,387	1,514	0,130
leagues.Bulgarian_u19	-1,979	1,274	-1,554	0,120
leagues.Club_Friendlies	2,017	1,117	1,806	0,070
leagues.Czech_Cup	18,769	1695,779	0,011	0,991
leagues.Greece_Cup	18,330	2399,545	0,008	0,993
leagues.Indonesia_Lower	-2,710	1,059	-2,558	0,010

leagues.Lithuania_A	17,514	2399,545	0,007	0,994
leagues.Russia_2_League	-2.172	0,950	-2,288	0,022
leagues.Slovenia_Prva_Liga	17,921	2399,545	0,007	0,994
leagues.Uzbekistan_B	-18,464	2399,545	-0,008	0,994

Tabela 12. Podsumowanie Modelu 3.

Źródło: Opracowanie własne (Rstudio)

Metoda krokowa wsteczna dokonywała wyboru najlepszego podzbioru, biorąc pod uwagę, jak najmniejszą wartość kryterium AIC. Pomimo znaczącej nieistotności wielu zmiennych, model ten pozwolił uzyskać poziom 75% trafności reklasyfikacji. Rozważając w kategorii szans, model pozwolił uzyskać trzykrotnie wyższą szansę na poprawną klasyfikację niż w przypadku losowym. Problem nieistotności mógł wynikać z niskiej liczby obserwacji w obrębie lig, których dotyczą problemowe zmienne. O niskiej mocy testu Walda w przypadkach liniowej separowalności klas wspomina także (Koronacki i Ćwik, 2005, s. 64-65).

Nazwa zmiennej	cup	event_level	drop	Afc_Cup	Azerbeijan_First
e^{β}	0,236	141,502	11,341	5,617	0,141
Nazwa zmiennej	Bosnia_Premier	Bulgaria_u19	Club_Friendlies	Czech_Cup	Greece_Cup
e^{β}	8,158	0,138	7,513	141595970,141	91396697,373
Nazwa zmiennej	Indonesia_Lower	Lithuania_A	Russia_2	Slovenia_Prva	Uzbekistan_B
e^{β}	0,066	40360719,762	0,113	60689953,515	0,001

Tabela 13. Wartości eksponentów współczynników pochodzących z modelu logitowego.

Źródło: Obliczenia własne

Mówiąc o wkładzie poszczególnych zmiennych na logarytm szans, spośród zmiennych ilościowych, **event_level** oraz **drop** wykazują dodatni wpływ, co oznacza, że zwiększanie się ich wartości, *ceteris paribus*, zwiększa szansę na trafne przewidzenie kierunku manipulacji. Na potrzeby dokładniejszej interpretacji oszacowano wartości eksponentów wszystkich zmiennych z Modelu 3 (Tabela 13). Pierwszy z wymienionych predyktorów, w przypadku wzrostu o 0,1 jednostki (w praktyce oznacza to występowanie w spotkaniu drużyn z niższych szczebli rozgrywek), w badanym modelu, zwiększa szansę na wystąpienie manipulacji 1,64 razy. Dla zmiennej **drop,** przyrost jej wartości o jednostkę, oznacza niemalże 12-krotnie większą szansę

na wystąpienie nadużyć. Dodatkowo fakt odbywania się spotkania w ramach Afc Cup, względem pozostałych rozgrywek, przynosi 5-krotnie wyższą szansę na manipulacje, dla ekstraklasy z Bośni, mnożnik ten oscyluje w okolicach wartości 8.Jeżeli chodzi o spotkania towarzyskie (**Club_Friendlies**), model sugeruje o około 650% większą szansę na potwierdzenia się kierunku podejrzanych ruchów kursowych w meczu, jeśli wartość tej zmiennej wynosi 1. Odwrotna sytuacja ma miejsce w przypadku zmiennej informującej o odbywaniu się spotkania w ramach pucharu (**cup**), w tym wypadku czynnik ten zmniejszałby o około 76% szansę na możliwość wystąpienia manipulacji. Pozostałe zmienne binarne, prawdopodobnie ze względu na jednostronne ukierunkowanie zmiennej objaśnianej w ich obrębie, wykazały bardzo ograniczający albo wyraźnie wysoki wpływ na szanse przewidzenia kierunku ruchów kursów.

Ostatni z modeli, wykazał zdecydowanie najniższą wartość kryterium AIC, wyróżnił się również, patrząc na trafność. Należy zauważyć, iż w tym wypadku, aby określić wyższość jednego z modeli, warto byłoby wziąć pod uwagę możliwość zastosowania predykcji dla nowych obserwacji. Przykładowo, próbując usystematyzować spotkanie, które odbyło się w ramach ligi, której do tej pory obserwacje nie dotyczyły, jedyną możliwością jest użycie Modelu 2, który klasyfikując mecz do jednej z dwóch grup, nie potrzebuje informacji o przynależności do konkretnych rozgrywek. Z tego powodu, pomimo teoretycznie gorszej ocenie, w kategoriach użytych miar, można wysnuć tezę o jego większej uniwersalności zastosowania. Oczywiście, mając do czynienia z meczem ligi, która do tej pory w obserwacjach wystąpiła, zalecałoby się użycie modelu ostatniego.

3.3 Zestawienie modeli

W wyniku działania obu metod, uzyskano modele o zbliżonej trafności, jeżeli chodzi o odwzorowanie próby uczącej. Modele uwzględniające zmienne dotyczące lig, w obu przypadkach składają się z analogicznego podzbioru predyktorów. Występują drobne różnice ze względu na konieczność pozbycia się pewnych zmiennych ligowych w analizie dyskryminacyjnej (zmienne stałe w grupach).

Postanowiono sprawdzić, jak zachowają się predykcje, bazujące na ostatecznych modelach otrzymanych w każdej z metod. W tym celu wylosowano obserwacje ze zbioru podstawowego.

liga	cup	event_level	alert	drop	Prawdopodobieństwo (Y=1)		result
					LDA	Regresja logistyczna	
Nikuragua_U20	0	0,65	0	0,35	0,754	0,781	1
AFC CUP	1	0,67	0	0,10	0,638	0,741	1
Club Friendlies	1	1,00	1	0,35	0,937	0,973	1
Vietnam_U19	0	0,65	0	0,35	0,797	0,781	1
Czech Cup	1	0,25	0	0,30	0,456	0,99	1
Nikaragua_1	0	0,33	0	0,04	0,493	0,258	0
Club Friendlies	1	0,40	0	0,20	0,634	0,563	1

Tabela 14. Prawdopodobieństwa klasyfikacji do grupy "1", dla obu metod. Obserwacje zostały wybrane w sposób losowy. Szarym kolorem oznaczono błędne klasyfikacje.
Źródło: Obliczenia własne

Modele uzyskane przy zastosowaniu zarówno regresji logistycznej, jak i analizy dyskryminacyjnej dla przedstawionych obserwacji, na ogół, wykazały zbliżone poziomy prawdopodobieństw klasyfikacji do grupy „1", co widoczne jest w Tabeli 14. Różnice wykreowały się dla spotkań z czeskiego pucharu, gdzie miał też miejsce jedyny błędny przypadek klasyfikacji (w obrębie wylosowanej podpróby) oraz pierwszego szczebla rozgrywek w Nikaragui.

Pierwsza z wymienionych obserwacji, zróżnicowała modele w stopniu znacznym, w tym wypadku na korzyść metody regresji logistycznej, która określiła prawdopodobieństwo przyporządkowania spotkania do grupy, w której wynik spotkania był zgodny z kierunkiem obniżających się kursów, z niemalże 100% pewnością.

Należy jednak zaznaczyć, iż sytuacja ta, wynika z faktu obecności w modelu utworzonego przy pomocy szczególnego przypadku regresji, zmiennej bezpośrednią związanej z rozgrywkami Czech Cup. W takim razie, na tej podstawie nie można wnioskować o przewadze tej metody, jeśli chodzi o modelowanie podejrzanych spotkań. Z drugiej strony, zastosowanie LDA w tym wypadku implikowało błędną klasyfikację. Analiza obserwacji dotyczącej spotkania z kraju położonego w Ameryce Środkowej, doprowadziła do odnotowania nieco mniejszej różnicy, jednakże modelowanie przy użyciu podejścia dyskryminacyjnego wskazało tu na bardzo bliską progowej (duża szansa na błędną klasyfikacje przy analogicznych obserwacjach), w zestawieniu z przypisaniem obserwacji do grupy „0" z dużą pewnością w przypadku drugiej metody. W ocenie autora, po uwzględnieniu dodatkowo, bardzo niskiej wartości predyktora **drop** oraz szczebla rozgrywek, pokazuje to przewagę modelu logitowego.

W celu otrzymania pełniejszego porównaniu, modele poddano testowi, polegającemu na klasyfikacji czterech nowych obserwacji, spoza zbioru uczącego (Tabela 15).

Liga	cup	event_level	alert	drop	Prawdopodobieństwo (Y=1)		result
					LDA	Regresja logistyczna	
Vietnam_u19	0	0,65	0	0,75	0,86	0,783	1
Vietnam_u19	0	0,65	0	0,25	0,77	0,68	0
Czech 2 League	1	0,25	0	0,15	0,48	0,38	0
Club Friendlies	1	0,5	1	0,12	0,66	0,54	1

Tabela 15. Klasyfikacja obserwacji spoza zbioru uczącego, przez modele utworzone przy pomocy LDA i regresji logistycznej. Błędne predykcje oznaczono kolorem szarym.

Źródło: Obliczenia własne

Również w tym wypadku, modele również wykazywały tendencję do szacowania prawdopodobieństw na podobnym poziomie. Prognozy, w obu przypadkach, okazały się nieprawidłowe dla umieszonej w drugim wierszu tabeli obserwacji, a trafne w pozostałych. Wracając do błędnie sklasyfikowanego spotkania, z tej perspektywy, idealny model musiałby uwzględniać w takiej sytuacji, w większym stopniu zmienną **drop.** Widoczne jest to, w momencie zestawienia z obserwacją pierwszą, analogiczną, oba modele oszacowały tutaj dość wysokie szanse przydzielenia do grupy „1", aczkolwiek, znaczny spadek wartości zmiennej drop, nie obniżył prawdopodobieństw wystarczająco.

Warto także nadmienić, iż w zbiorze predyktorów, składających się na model wywodzący się z LDA, znajduje się zmienna binarna dotycząca wietnamskiej ligi młodzików, co powinno pozwolić na trafniejszą w tym przypadku, w porównaniu z modelem logitowym, klasyfikację.

Podsumowując, w podejściu ogólnym oba modele zachowywały się niezwykle podobnie. Zagłębiając się jednak, w przypadki szczególne, wyklarowała się nieznaczna przewaga modelowania podejrzanych spotkań, przy pomocy regresji logistycznej. W dodatku, nasuwa się teza, o uwzględnieniu w zbyt małym stopniu predyktora dotyczącego wartości obniżających się kursów, w obu modelach.

3.4 Predykcja

Dzięki możliwości oszacowania prawdopodobieństwa klasyfikacji obserwacji do poszczególnych grup, w połączeniu ze znanym (przytoczonym w części wstępnej)

sposobem jego konwersji na dziesiętne kursy bukmacherskie, zdecydowano zaproponować, prosty system doboru stawek do zakładów, które hipotetyczny inwestor mógłby zawierać, obserwując spotkania, w których widoczne są duże wahania kursów.

Na wstępie, należałoby rozgraniczyć, na jakim progu, zapada decyzja o wstrzymaniu się od zawarcia zakładu. Badacz sugeruje określeniu tego poziomu na 0,5. Wartości oszacowanego prawdopodobieństwa podzielono na 10 przedziałów, a następnie przełożono na sugerowaną stawkę, w skali 1-10, gdzie nota najwyższa oznacza maksymalną stawkę. Proponowany system postanowiono zaprezentować dla obserwacji z próby uczącej (Tabela 16). Do estymacji wybrano model logitowy. Ze względu na brak wystarczająco dokładnych danych dla wszystkich obserwacji, przyjęto założeniu o możliwości zawarciu zakładu, w każdym przypadku po kursie 2,0 (w zdecydowanej większości obserwacji, dostępne kursy, w momencie pojawienia się różnic kursowych, przekraczały ten poziom, najniższa wartość kursu, jaka wystąpiła to 1,9).

Obliczony został również hipotetyczny zysk, możliwy do osiągnięcia przez potencjalnego inwestora, który zdecydowałby się podejmować decyzje oparte o działanie zaproponowanego systemu. Należałoby założyć, iż taka osoba, zawarłaby zakłady dotyczące rezultatów spotkań, zgodnie z przewidywanym przez model kierunkiem dla wszystkich obserwacji z próby uczącej. Zakładając dodatkowo, że podejmowanie decyzji, co do wysokości stawek, odbywałoby tak jak sugeruje to model (Tabela 16), zsumowanie bilansu uzyskanego ze wszystkich spotkań, pozwoliłoby wygenerować profit w postaci 483 jednostek. Dodatkowo, na Rysunku 12 widoczne jest, w jaki sposób kształtowałaby się wartość bilansu portfela, wraz ze uwzględnianiem kolejnych spotkań. Warto dodać, iż bilans ten, dla obserwacji z próby uczącej osiągał coraz wyższe wartości, nigdy nie zbliżyłby się też do poziomu ujemnego. W praktyce dla potencjalnego inwestora oznaczałoby to znaczne zmniejszenie ryzyka bankructwa.

Liga	cup	event_level	alert	drop	result	Stawka(j)	Zysk(j)	P (Y=1)
Czech U19 Legaue	0	0,65	0	0,45	0	7	-7	0,82
Greece Football League	0	0,60	0	0,35	1	5	5	0,74
Vietnam Second Division	0	0,75	0	0,45	1	8	8	0,88
Club Friendlies	1	0,75	0	0,15	1	8	8	0,87
Greece Football League	0	0,60	0	0,40	1	6	6	0,76
Vietnam Second Division	0	0,75	0	0,50	1	8	8	0,89
Vietnam U19 League	0	0,65	0	0,35	1	6	6	0,78
Nikuragua Primiera	0	0,33	0	0,04	0	0	0	0,26
Nikuragua U20 Matches	0	0,65	0	0,15	0	4	-4	0,69
Greece Football League	0	0,60	1	0,25	1	4	4	0,69

Tabela 16. Dobór stawek dla wybranych obserwacji ze zbioru wstępnego.

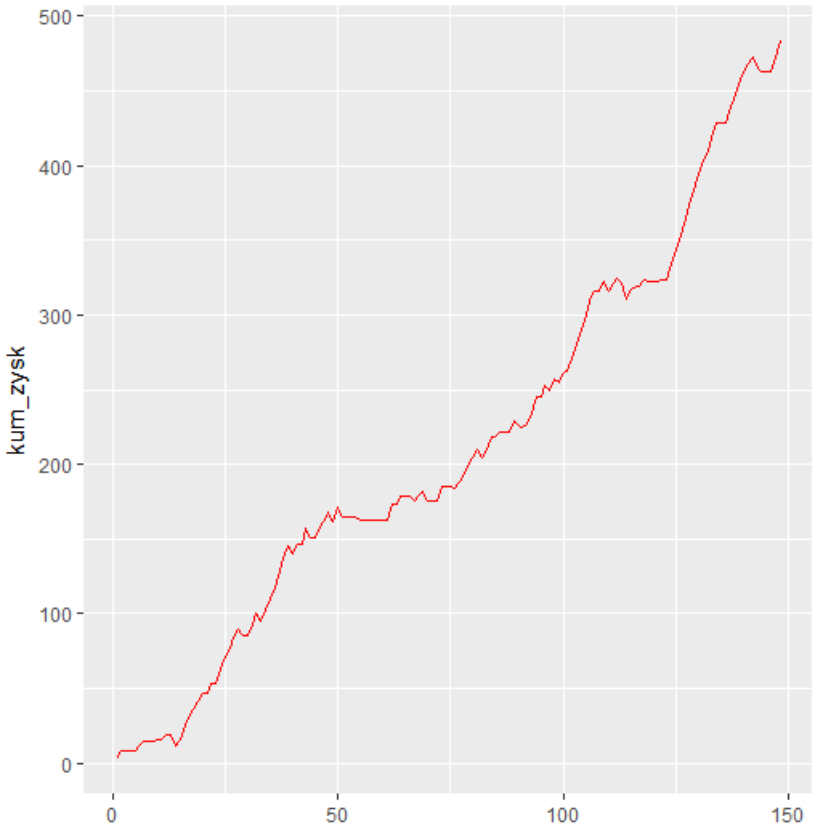
Źródło: Obliczenia własne

Patrząc z innej perspektywy, mając do dyspozycji modele i oszacowane na ich podstawie prawdopodobieństwa, badacz proponuje system wspomagania, kontroli kursów dla podmiotów będących stroną przyjmującą zakłady (Tabela 17). W tym wypadku, prawdopodobieństwa zamieniano by wprost na ofertę kursową, dopasowaną do tego szczególnego typu spotkań, w których dochodzi do nietypowych wahań. Podejście takie umożliwiłoby, uwzględniając marże, zakładając odpowiedni dobór modelu, w długim terminie, uzyskanie przewagi nad graczami, dzięki posiadaniu oferty kursowej, która prawidłowo oddaje prawdopodobieństwa. Zasugerowany wariant, na potrzeby prezentacji wyników, uproszczono do przypadku rynków dwudrogowych.

liga	cup	event_level	alert	drop	result	P (Y=1)	kurs_1	kurs_0
Czech U19 Legaue	0	0,65	0	0,45	0	0,82	1,22	5,56
Greece Football League	0	0,60	0	0,35	1	0,74	1,36	3,81
Nikuragua Primiera	0	0,33	0	0,04	0	0,26	3,86	1,35

Tabela 17. Oferta kursów bukmacherskich, bez uwzględnienia marży, stworzona na podstawie prawdopodobieństw pochodzących z modelu logitowego (wariant 2-drogowy).

Źródło: Obliczenia własne



Rysunek 12. Wykres skumulowanej sumy hipotetycznego zysku z inwestycji, zakładając używanie zaproponowanego modelu do doboru stawek. Oś pozioma przedstawia zakres obserwacji, a więc kum_zysk(x) przedstawia wartość bilansu portfela w momencie, którym dokonano już hipotetycznego zawarcia zakładów na spotkania z indeksem od nr 0 do x.

Źródło: Opracowanie własne (Rstudio)

4. Podsumowanie

Praca miała na celu opracowanie modelu wspomagającego detekcję meczów, w których może dochodzić do manipulacji. Okazało się, iż możliwa jest także ocena prawdopodobieństwa klasyfikacji takich spotkań do grupy podejrzanych, pod względem możliwości wystąpienia nadużyć. Pozwoliło to znaleźć dla modelów zastosowania praktyczne. Zaproponowano system wspomagania decyzji inwestycyjnych dla potencjalnych inwestorów, jak i również algorytm pomocny, jeśli chodzi o kontrolę, czy też tworzenie oferty kursowej.

Ważną kwestią okazało się również zbadanie czynników, które na możliwość przewidzenia kierunku nadużycia, miały największy wpływ. Po przeprowadzeniu badań, śmiało można stwierdzić, że poziom, na jakim toczy się mecz (określany przez zmienną **event_level)**, ewidentnie do takich czynników się zalicza. Predyktor ten, obecny był w niemalże każdym rozważanym modelu. Praktycznie wszystkie użyte miary pokazywały jego istotność statystyczną. Na poprawę skuteczności modeli duży wpływ miało także dodawanie zmiennych związanych z konkretnymi rozgrywkami, do których przynależały spotkania. Pokazuje to, iż występują pewne wzorce, jeśli chodzi o szanse na manipulacje, w obrębie poszczególnych lig. Idąc dalej tym tropem, nasuwa się idea tworzenia większej ilości modeli, nieco mniej uniwersalnych, za to wyraźnie skuteczniejszych, np. w obrębie danych rejonów geograficznych.

Wracając do wstępnego zbioru predyktorów, badania sugerują, iż zmienna **game_progress** nie miała większego znaczenia dla poprawy jakości modeli. Nie znalazła ona w również w żadnym z końcowych modeli. Próby modelowania rezultatów spotkań z jej uwzględnieniem wydają się nieskuteczne. Co ciekawe, przeczy temu raport (2019), widoczna jest tam prawidłowość, pokazująca, że spora część podejrzanych odbywa się w ramach ostatnich faz rozgrywek. Wytłumaczeniem takiej rozbieżności może być fakt, iż w publikowanych raportach nie uwzględnia się kierunku manipulacji. Autorzy tych publikacji skupiają się jedynie na wychwyceniu spotkań, w których zawierane są niepokojące zakłady. Z kolei w przypadku tej pracy, podejmuje się także próbę przewidzenia, jak dokładnie zmanipulowane spotkanie będzie przebiegać (jakim rezultatem się zakończy, kto zdobędzie następną bramkę itp.). W takiej sytuacji, dane z raportów publikowanych przez takie organizacje można by porównywać do zebranego zbioru obserwacji.

Porównując analizę dyskryminacyjnej oraz regresję logistyczną w kontekście badań nad omawianą tematyką, autor zauważa niewielką przewagę drugiej ze wspominanych technik. Model logitowy pozwolił uzyskać minimalnie wyższy odsetek trafności klasyfikacji. W dodatku, stosując regresję logistyczną, badacz nie jest ograniczany przez liczne założenia.

Warto jeszcze raz podkreślić, iż modelowanie spotkań podejrzanych o zachodzenie manipulacji może okazać się przydatnym narzędziem z punktu widzenia różnych podmiotów. Pokazano, iż możliwe jest szacowanie prawdopodobieństwa, a co za tym idzie tworzenie zarówno oferty kursowej, jak i wypracowanie długoterminowej taktyki związanej z zawieraniem zakładów w celu generowania zysków. Posiadanie odpowiedniego modelu związanego z badanym tematem, z pewnością, poza czysto technicznym podejściem, próbą wykorzystania matematyki do osiągania korzyści finansowych, mogłoby ułatwiać detekcję i ukierunkowywać odpowiednie działania mające zapobiegać manipulacjom w sporcie.

Aby modelowanie mogło stać się skuteczniejsze, dobrze byłoby posiadać znaczącą ilość obserwacji z obrębu każdej z lig. Dodatkowo, należałoby stworzyć narzędzie analizujące podejrzane ruchy w trybie ciągłym, co pozwoliłoby zebrać nieporównywalnie większą liczbę danych. Możliwość analizy obszernej ilości predyktorów z pewnością wpłynęłoby pozytywnie na jakość otrzymanego modelu. Z kolei faktem negatywnie wpływającym na istotność niektórych zmiennych, a zarazem na pogarszającym wyniki estymacji, była spora liczba braków danych, rozszerzając badania, autor sugeruje podjęcie próby uzupełnienia tychże braków poprzez zastosowanie efektywniejszych metod.

Bibliografia

- Bruce, P., Bruce A., (2017). *Practical Statistics for Data Scientists: 50 Essential Concepts*, O'Reilly Media
- Cortis, D. (2015). *Expected values and variances in bookmaker payouts: A theoretical approach towards setting limits on odds*, The Journal of Prediction
- Danieluk B. (2010). *Zastosowanie regresji logistycznej w badaniach eksperymentalnych*, Psychologia Społeczna, tom 5 2–3 (14) 199–216, Instytut Psychologii UMCS, Lublin
- Hosmer, D. W., Lemeshow, S. (2000). *Applied logistic regression*, wyd. 2, Wiley & Sons, Nowy Jork
- Koronacki, J., Ćwik, J. (2005). *Statystyczne systemy uczące się*, Exit, Warszawa
- Markets, 9(1):1, University of Leicester, University of Malta
- Prabhakaran, S., (2017). *Logistic Regression – A Complete Tutorial With Examples in R*, <https://www.machinelearningplus.com/machine-learning/logistic-regression-tutorial-examples-r> [dostęp 09.06.2020]
- Radkiewicz P. (2010). *Analiza dyskryminacyjna. Podstawowe założenia i zastosowania w badaniach społecznych*, Psychologia Społeczna, tom 5 2-3 (14) 142-161, Instytut Studiów Społecznych, Uniwersytet Warszawski
- Stats Perform, (2019). *Suspicious Betting Trends in Global Football 2019*, <http://docs.statsperform.com/Publications/Suspicious_Betting_Trends_2019.pdf>
- Zhongheng, Z., (2016). *Model building strategy for logistic regression: purposeful selection*, Ann Transl Med, 4(6): 111

Spis tabel

Spis rysunków