Analiza i wizualizacja danych

Projekt dotyczy zakładów sportowych. Przewidywania, czy zespół gospodarzy wygra. Czy jesteśmy w stanie stworzyć model zdolny do dokładnego przewidywania czy zespół gospodarzy wygra mecz? Który wytrenowany model będzie bardziej skuteczny?

Analizowany zbiór danych zawiera wyniki z niemieckiej Bundesligi z okresu 1993-2018. Zbiór został wzięty ze strony:

<https://www.kaggle.com/thefc17/bundesliga-results-19932018/version/1>

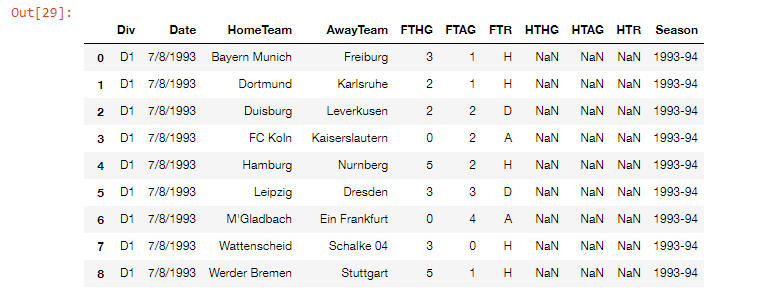
który z kolei podejrzewam, że został uzyskany przez złączenie i wyczyszczenie informacji o wszystkich sezonach w jeden plik na podstawie strony:

http://www.football-data.co.uk/germanym.php

Zbiór zawiera informacje takie jak: drużyna gospodarzy, drużyna gości, gole strzelone u siebie, gole strzelone na wyjeździe, wynik końcowy (H-zwycięstwo gospodarzy, A-zwycięstwo gości, D-remis), gole strzelone przez gospodarzy do przerwy, gole strzelone przez gości do przerwy, wynik do przerwy (H-przewaga gospodarzy, A-przewaga gości, D-remis), data i sezon, w którym dany mecz był rozegrany.

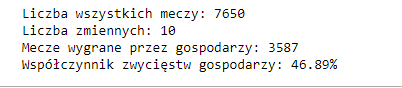
Analiza zbioru danych:

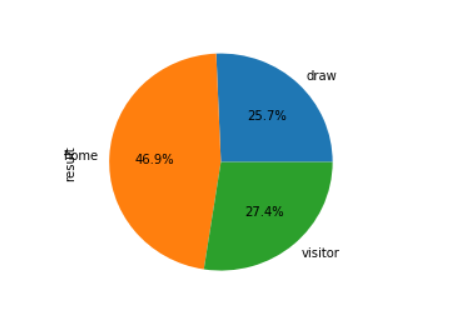
Dla pierwszych dwóch sezonów nie zostały zachowane/wprowadzone kolumny HTHG (gole strzelone przez gospodarzy do przerwy), HTAG (gole strzelone przez gości do przerwy), HTR (wynik do przerwy). W powyższych kolumnach widzimy NaN, co oznacza not a number. Dla pozostałych sezonów wartości dla tych kolumn są już uzupełnione wartościami, co można zobaczyć wyświetlając większą liczbę rzędów w zbiorze.





W profesjonalnej piłce nożnej często za przewagę uznaje się grę na własnym stadionie. Np. w rozgrywkach Ligi Mistrzów w fazie pucharowej w momencie takiego samego bilansu bramkowego po obu meczach, do dalszego etapu awansowuje drużyna, która strzeliła więcej bramek na wyjeździe. Z tego względu warto zobaczyć jak kształtuje się bilans wyników drużyn grających w Bundeslidze przyjmując kryterium gry na własnym stadionie.





Jak widzimy do typowania zwycięzcy danego spotkania w lidze niemieckiej warto wziąć pod uwagę kryterium tego, czy mecz odbywa się na stadionie gospodarza. Można by zastanawiać się skąd faktycznie bierze się przewaga własnego boiska. Czy wynika to z obecności większej liczby kibiców zespołu gospodarzy, z tego czy piłkarze nie muszą specjalnie podróżować na mecz, lepiej znają własny stadion? Ja osobiście uważam, że największy wpływ może mieć faworyzowanie drużyny gospodarzy przez sędziego na skutek presji ze strony kibiców. Sędziowie podświadomie mogą dokonywać przychylniejszych decyzji dla gospodarzy, żeby nie narazić się na atak ze strony nieprzychylnego tłumu.

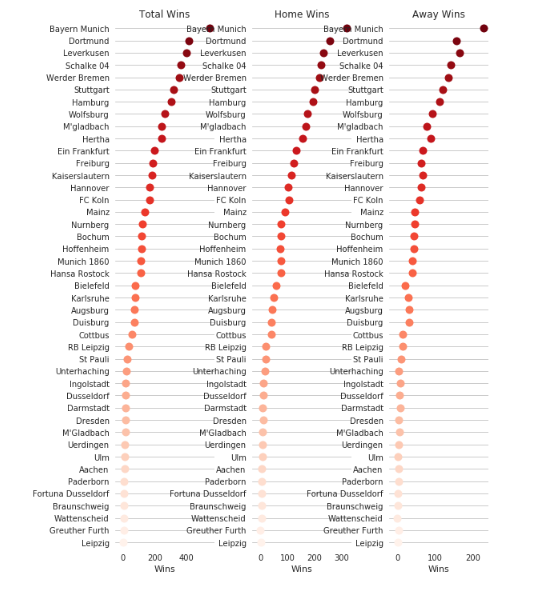
Które zespoły wygrały średnio najwięcej gier u siebie?



Które zespóły wygrały średnio najwięcej gier na wyjeździe?



Jak widać 1 ligę niemiecką od lat dominuje Bayern Monachium. W obu wyliczonych średnich znalazły się kilka takich samych drużyn, które na przestrzeni lat rozegrały setki meczów w lidze. To zespoły takie jak Leverkusen, Dortmund, Schalke 04, Werder Bremen, Stuttgart. Może to świadczyć o dominacji tej ligi przez te zespoły i prognozować dobre wyniki tych drużyn. W zestawieniu z wysokimi średnimi zwycięstw znalazły się także pojedyncze zespoły, które rozegrały stosunkowo mało meczy w 1 lidze. Zespół RB Leipzig powstał dopiero w połowie 2009 roku. A debiut w najwyższej lidze zanotował 28 sierpnia 2016. M'Gladbach w latach 1999-2011 przeżywał kryzys, podczas którego spadał do 2. Bundesligi. 2011 klub powrócił do ligowej czołówki, co miało wpływ na pojawienie się w zestawieniu ze średnimi zwycięstwami. Ostatni klub z małą liczbą meczy, który znalazł się tylko w jednym zestawieniu ze średnimi zwycięstwami (zwycięstwa u siebie) jest Unterhaching. Klub ten w sezonach 1999/2000 i 2000/2001 występował w Bundeslidze. Obecnie gra w 3 lidze, ale jak widzimy w tych dwóch sezonach w najwyższej lidze na swoim stadionie radził sobie całkiem nieźle.



Przejdźmy do trenowania modeli:

Jest to problem klasyfikacji. Przewidzenie, czy zespół gospodarzy wygra czy przegra.

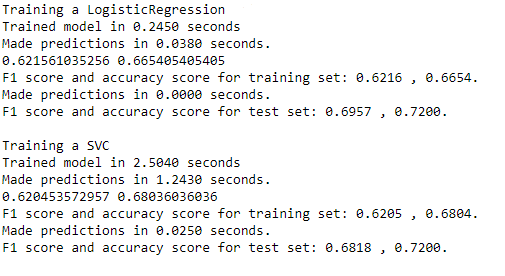
Użyjmy więc jednej z najpopularniejszych odmian analizy regresji -**regresji logistycznej (**określamy wystąpienie bądź brak wystąpienia pewnego zdarzenia**).**

**Do kolejnego modelu wykorzystamy metodę wektorów nośnych.**

Wytreujemy je na zbiorze danych. Wybierzemy najskutecznieszy klasyfikator z posród nich do przewidzenia, który zespół wygra. W tym celu użyjemy historii wyników.

Ustalamy zmienną celu jako FTR – full time result. To co chcemy zrobić to dać naszemu modelowi wszystkie pozostałe, dostępne cechy i na tej podstawie spróbować przewidzieć FTR.

W tym celu dzielę sobie zbiór na dane bez FTR i na zbiór FTR. Standaryzuję dane do takiej samej skali (podobny typ zmiennych, i skaluję- czyli unikamy sytuacji, w której jedna zmienna jest np. w setkach tysięcy, a inna w dziesiątkach, czyli jeżeli mamy gdzieś małe wartości, to chcemy żeby wszystkie one (wartości) były małe). Taki zabieg poprawia możliwości przewidywania naszego modelu.

****

**Wytrenowanie i przewidzenie przez metodę wektorów nośnych zajęło zdecydowanie dłużej. Obie metody okazały się być zbliżenie dokładne wg F1-score.**