POLITECHNIKA GDAŃSKA

Wydział Elektroniki, Telekomunikacji i Informatyki

Kierunek: Informatyka

Dobór rozmiaru latawca kitesurfingowego na podstawie wagi użytkownika i prędkości wiatru z użyciem metod regresji

Autor: Maciej Daszkiewicz

Spis treści

1 Wstęp

Kitesurfing to dynamicznie rozwijający się sport wodny, który polega na poruszaniu się po wodzie przy pomocy deski i latawca napędzanego wiatrem. Jednym z kluczowych elementów wpływających na komfort i bezpieczeństwo podczas uprawiania kitesurfingu jest prawidłowy dobór rozmiaru latawca do panujących warunków atmosferycznych i cech fizycznych użytkownika.

W praktyce najczęściej stosuje się empiryczne podejścia do doboru rozmiaru, opierające się na doświadczeniu, ogólnych tabelach lub prostych wzorach typu „waga podzielona przez wiatr razy współczynnik". Takie podejścia są wygodne, lecz nie zawsze precyzyjne – nie uwzględniają bardziej złożonych relacji pomiędzy zmiennymi, jak również różnorodności warunków.

Celem niniejszego projektu jest analiza tego problemu z użyciem metod uczenia maszynowego – a konkretnie regresji. Chcemy sprawdzić, czy możliwe jest uzyskanie dokładniejszego dopasowania rozmiaru latawca do warunków takich jak waga użytkownika i prędkość wiatru, bazując na rzeczywistych danych zebranych od kitesurferów.

1.1 Opis problemu

Z fizycznego punktu widzenia, latawiec musi wygenerować wystarczającą siłę ciągu, aby unieść kitesurfera z odpowiednią prędkością na wodzie. Jeśli jest zbyt mały względem warunków (słaby wiatr lub ciężki użytkownik), nie będzie w stanie go unieść; z kolei zbyt duży latawiec w silnym wietrze może generować niekontrolowane przeciążenia, stanowiąc zagrożenie dla zdrowia.

Aby uprościć analizę, założono że:

- deska to klasyczny twin-tip,
- latawiec to standardowy pompowany kite,
- W analizie pominięto inne czynniki (umiejętności, długość linek, typ wiatru).

Analizujemy więc zależność pomiędzy dwoma zmiennymi wejściowymi:

- waga użytkownika (kg),
- prędkość wiatru (węzły),

a oczekiwaną zmienną wyjściową:

- rozmiar użytego latawca (m^2).

2 Regresja liniowa - Teoria

Regresję liniową można sobie wyobrazić jako próbę dopasowania prostej (lub w ogólniejszym przypadku — płaszczyzny lub hiperpowierzchni), która możliwie najlepiej przechodzi przez zbiór punktów na wykresie. Każdy punkt odpowiada jednej obserwacji (np. jednemu

użytkownikowi z konkretną wagą i prędkością wiatru) i jest opisany zarówno przez dane wejściowe (cechy), jak i przez oczekiwaną wartość wyjściową (rozmiar latawca). Regresja liniowa stara się znaleźć równanie, które pozwala dla nowych danych (nowej osoby i nowej prędkości wiatru) podać jak najlepszą prognozę rozmiaru.

Najprostszy przypadek regresji można zapisać jako:

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

Gdzie:

- $\hat{y}$ — przewidywana wartość (rozmiar latawca),
- $x_1, x_2, \ldots$ — dane wejściowe (np. masa, wiatr, ich przekształcenia),
- $w_0, w_1, \ldots$ — tzw. wagi (parametry modelu), które algorytm dobiera podczas nauki.

2.1 Rozszerzenie: cechy nieliniowe

Chociaż model jest „liniowy" w sensie matematycznym (czyli nie zawiera np. pierwiastków z wag, logarytmów z wag itd.), to może być bardzo elastyczny, jeśli do środka wstawimy odpowiednie cechy.

W projekcie oprócz podstawowych cech:

- masa ciała,
- prędkość wiatru,

stworzono wiele przekształceń tych wartości, takich jak:

- iloczyn: $masa \cdot wiatr$,
- ilorazy: $masa/wiatr$, $masa/wiatr^2$, $wiatr/masa$, $wiatr/masa^2$,
- inne wybrane kombinacje.

Choć model pozostaje matematycznie liniowy, to dzięki tym nieliniowym przekształceniom może opisywać dużo bardziej złożone zależności.

2.2 Standaryzacja danych

Wszystkie dane wejściowe zostały poddane **standaryzacji**, czyli przekształcone tak, aby miały:

- średnią wartość równą 0,
- odchylenie standardowe równe 1.

Takie przekształcenie jest potrzebne, ponieważ różne cechy mogą mieć bardzo różne skale (np. masa liczona w kilogramach, a inne cechy to bardzo małe ułamki). Bez standaryzacji cechy o dużych wartościach mogłyby „zdominować" działanie algorytmu. Dodatkowo, standaryzacja ułatwia i przyspiesza proces uczenia się parametrów.

2.3 Mean Squared Error (MSE)

Jedną z podstawowych miar jakości modelu regresji jest *Mean Squared Error* (MSE), czyli średni błąd kwadratowy. MSE mierzy średnią kwadratową różnicę między wartościami rzeczywistymi y_i a przewidywanymi $\hat{y}_i$:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

MSE przyjmuje zawsze wartość nieujemną, a im jest mniejsza, tym lepsze dopasowanie modelu do danych. Ze względu na podnoszenie błędów do kwadratu, większe różnice są silniej karane, co czyni MSE wrażliwym na wartości odstające.

2.4 Uczenie modelu — spadek gradientu

Aby dobrać parametry $w_0, w_1, \ldots, w_n$, wykorzystano metodę optymalizacji zwaną **spadkiem gradientu** (ang. *gradient descent*).

Algorytm ten działa intuicyjnie podobnie do „schodzenia ze wzgórza": zaczynamy od losowego miejsca (losowych wartości wag), obliczamy błąd (czyli jak bardzo nasz model się myli), a następnie krok po kroku poprawiamy parametry, idąc w kierunku, który zmniejsza ten błąd.

Dla każdej aktualizacji parametrów wykorzystujemy wzór:

$$w_j \leftarrow w_j - \eta \cdot \frac{\partial L}{\partial w_j}$$

Gdzie:

- w_j — parametr, który aktualizujemy,
- η — tzw. współczynnik uczenia (jak duży krok robimy),
- $\frac{\partial L}{\partial w_j}$ — pochodna funkcji błędu względem tego parametru (czyli kierunek, w którym model powinien się poprawić).

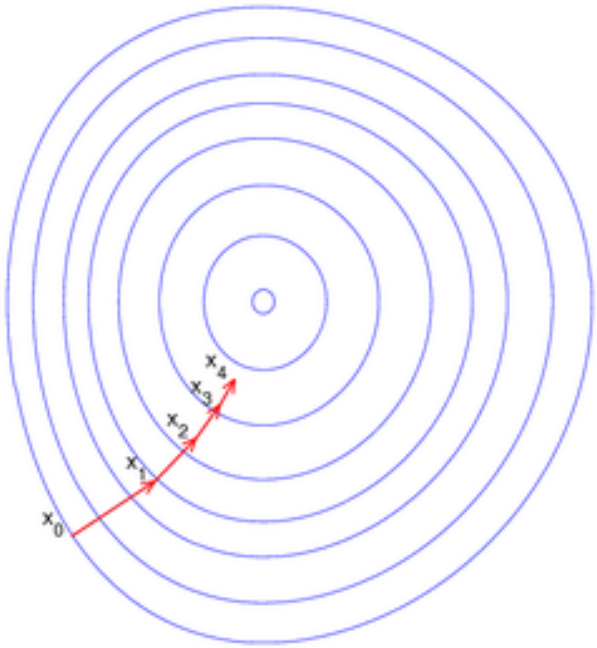
Proces ten jest powtarzany aż do momentu, gdy błąd przestaje się znacząco zmniejszać (czyli osiągniemy satysfakcjonujące dopasowanie).

3 Dane

3.1 Zebranie danych

Dane zostały zebrane za pomocą interaktywnej ankiety:

`https://mlodyd.github.io/Badanie-Rozmiar-Latawca/`



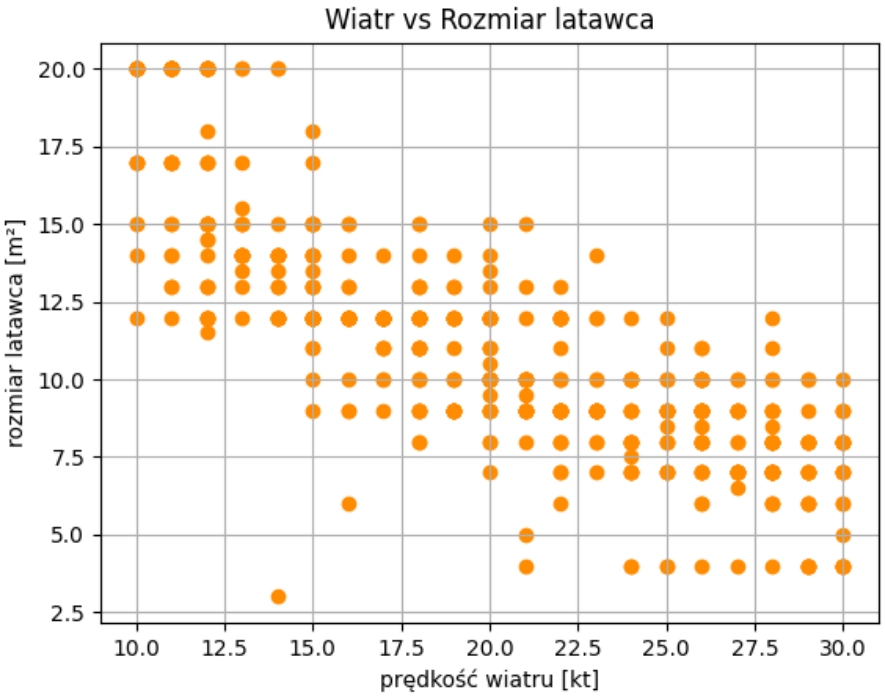
Rysunek 1: Źródło grafiki: `https://pl.wikipedia.org/wiki/Metoda_gradientu_prostego`

Ankieta była udostępniona na grupie „Kite Forum Polska" na portalu społecznościowym Facebook, co pozwoliło na zebranie ponad 500 rekordów. Każdy uczestnik podawał swoją wagę oraz rozmiary latawców, jakich używa przy losowo dobranych wartościach prędkości wiatru (10–30 węzłów).

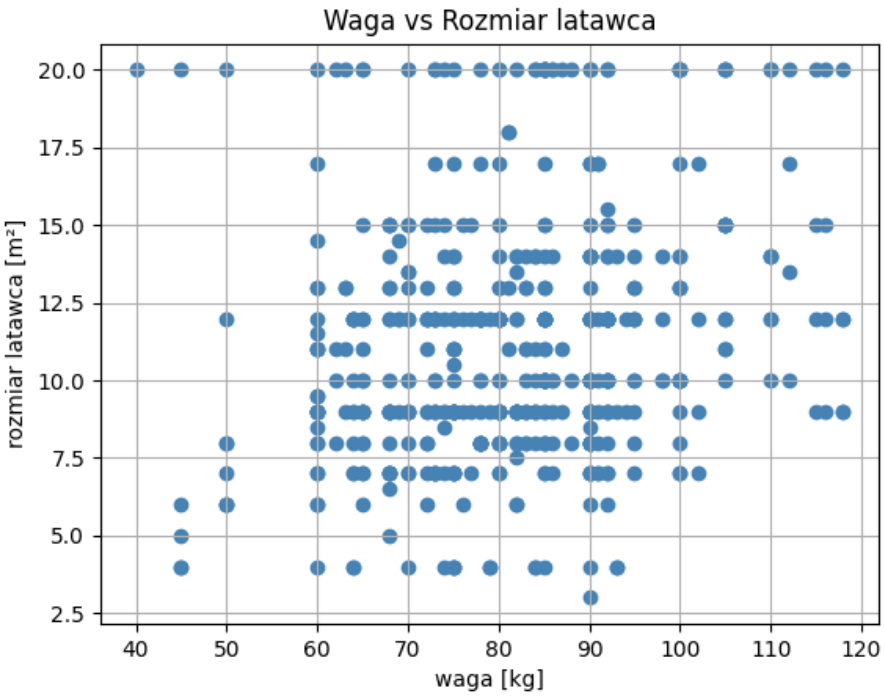
W przypadkach, gdy użytkownik stwierdził, że przy danej prędkości „nie da się pływać", przypisywano wartość 18 m^2, a gdy wiatr był „za mocny" – 5 m^2. Pozwoliło to uwzględnić również dane brzegowe w analizie.

3.2 Wstępna analiza danych

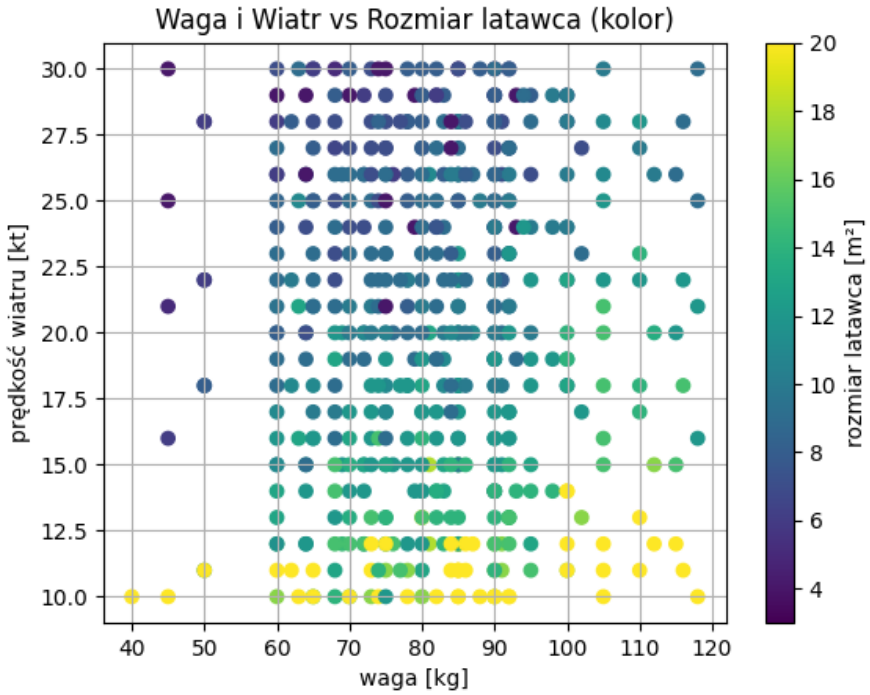
Po oczyszczeniu danych i przekształceniu odpowiedzi do formy liczbowej, wykonano analizę zależności między zmiennymi. Wykresy przedstawiające związek pomiędzy wagą, prędkością wiatru a rozmiarem latawca przedstawiam poniżej:



Rysunek 2: Zależność pomiędzy prędkością wiatru a rozmiarem latawca



Rysunek 3: Zależność pomiędzy wagą kitesurfera a rozmiarem latawca



Rysunek 4: Zależność pomiędzy prędkością wiatru i wagą a rozmiarem latawca

Obserwacje:
malejący rozmiar latawca wraz ze wzrostem prędkości wiatru,
rosnący rozmiar wraz ze wzrostem wagi użytkownika.

Wizualizacja trójwymiarowa *(waga, wiatr, rozmiar)* ukazała wygładzoną powierzchnię zależności, co sugeruje, że metody regresyjne będą odpowiednie.

4 Modelowanie

Dane dzielono losowo na zbiór treningowy i testowy (80/20). Wykorzystano własną implementację regresji liniowej trenowaną z użyciem spadku gradientu prostego oraz funkcji kosztu MSE (średni błąd kwadratowy). Każdą kolumnę w macierzy obserwacji ustandaryzowano ze średnią równą zero oraz odchyleniem standardowym równym jeden. Poza liniowymi czynnikami do macierzy obserwacji dodane zostały również nieliniowe kombinacje prędkości wiatru i wagi

5 Transformacje cech

Aby uchwycić nieliniowe zależności, przygotowano 13 funkcji opartych na dwóch zmiennych:

0) funkcja stała,

1) *waga*,

2) $predkosc\,wiatru$,

3) $\frac{waga}{predkosc\,wiatru}$

4) $\frac{predkosc\,wiatru}{waga}$

5) $waga \cdot predkosc\,wiatru$

6) $waga^2$,

7) $prdko\,wiatru^2$,

8) $\frac{waga^2}{predkosc\,wiatru}$

9) $\frac{predkosc\,wiatru^2}{waga}$

10) $\frac{predkosc\,wiatru}{waga^2}$

11) $\frac{waga}{predkosc\,wiatru^2}$

12) $waga^3$

13) $predkosc\,wiatru^3$

Dla każdej kombinacji maksymalnie 5 cech modele trenowano i mierzono ich błędy na całym zbiorze danych, aby upewnić się, że dane cechy są w stanie dobrze dopasować się do zbioru treningowego.

6 Wyniki i analiza

Najlepsze wyniki dla 2 cech:

- użyto funkcji: [0,3], a wartość MSE wynosiła 3.33
- użyto funkcji: [0,11], a wartość MSE wynosiła 3.31

Najlepsze wyniki dla 3 cech:

- użyto funkcji: [0,1,3], a wartość MSE wynosiła 2.74
- użyto funkcji: [0,9,11], a wartość MSE wynosiła 2.70

Dla większej ilości cech wartość MSE nieznacznie się zmniejszała, więc mając na uwagę wartość prostoty rozwiązania postanowiłem je pominąć

Wybrana przeze mnie funkcja regresji po odstandaryzowaniu:

$$Rozmiar = 10{,}28 - 0{,}38 \cdot \frac{predkosc\,wiatru^2}{waga} + 10{,}05 \cdot \frac{waga}{predkosc\,wiatru^2}$$

Następnie przeprowadzono trening i test na podzielonych danych. Aby zminimalizować losowość i potencjalne szczęśliwe ułożenie danych przeprowadzono 1000 losowań podziału danych na dane treningowe i testowe oraz wykonano dla nich trening i obliczone zostały MSE dla danych testowych. Średnia wartość MSE wynosiła 2.74.

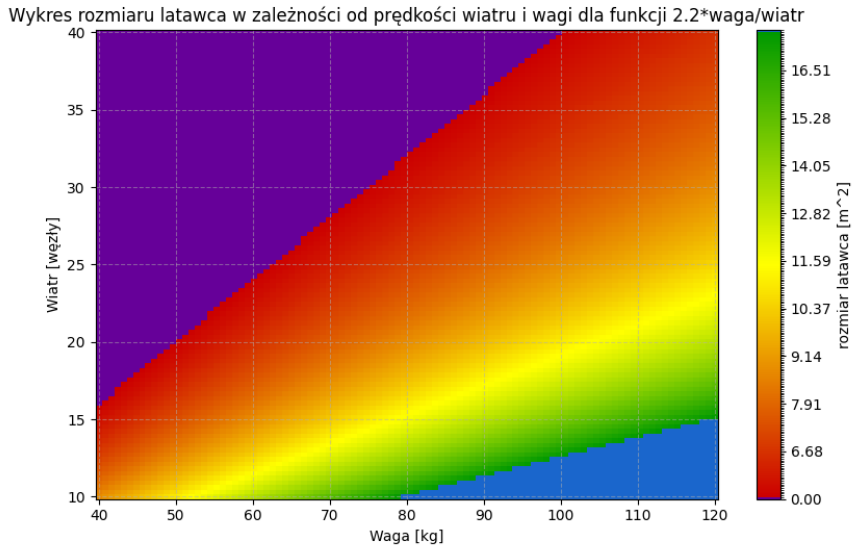
6.1 Porównanie z przyjętym wzorem

Popularny wzór wśród kitesurferów:

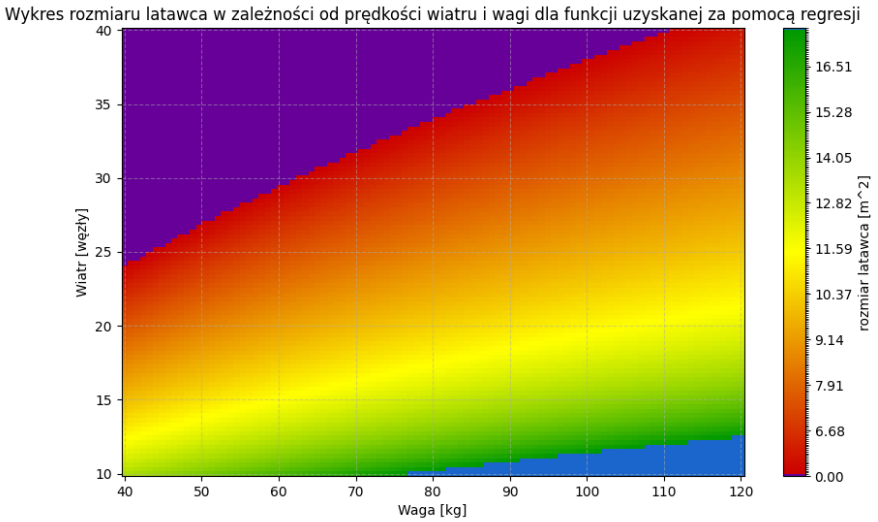
$$Rozmiar = 2{,}2 \cdot \frac{waga}{predko\,wiatru}$$

Okazał się mniej dokładny — jego MSE wyniosło około 4.75, a wyćwiczonego modelu wynosiło 2.65. Model regresyjny daje znacznie lepsze dopasowanie, jednak trzeba mieć na uwadze, że nie znaczy to jednoznacznie, że jest on lepszy, a jedynie mamy pewność, że jest on lepiej dopasowany do tych danych. Istnieje możliwość, że dla innych danych wyniki mogłyby być podobne albo nawet gorsze.

Aby móc jakoś porównać te modele, niezależnie od danych wykonano wykres wartości zwracanych przez model i przez wzór znany wśród kitesurferów.



Rysunek 5: Wykres dla wzoru znanego wśród kitesurferów



Rysunek 6: Wykres dla funkcji wyznaczonej przez regresję

Na wykresach fioletowym kolorem zostały oznaczone obszary w których wiatr jest za duży (rozmiar latawca jest mniejszy od $5.5m^2$), analogicznie niebieskim kolorem oznaczono obszar w którym wiatr jest za mały (rozmiar latawca jest większy od $17.5m^2$). Można zauważyć, że wykresy te są dosyć podobne, w obu wykresach zauważalny jest wzrost wartości w podobnym kierunku. Zakres wiatrowy według znanego wzoru jest mniejszy. Dla dużych prędkości wiatru można zaobserwować, że model regresyjny zachowuje się lepiej, ale dla małych wag kitesurfera działa zdecydowanie gorzej. Może być to spowodowane tym, że ilość danych w przedziale wagowym (40,60) kg była bardzo niewielka, bo poniżej 2% wszystkich danych

6.2 Podsumowanie, Wnioski i myśli okiem „Eksperta"

Dla modelu wytrenowanego regresją wartość MSE na danych wyniosła 2.65, a model średnio mylił się o $1.25m^2$. Nie jest to zły wynik, zważając na to, że przy jednej prędkości wiatru zazwyczaj (pomijając skrajnie silne i słabe wiatry) można pływać na latawcach, których rozmiary różnią się nawet o 3-4 metry kwadratowe. Mało kto posiada latawce, których rozmiary różnią się o metr. Zwykle kitesurferzy posiadają latawce różniące się o 2-3 metry kwadratowe, więc takie lekko niedokładne przewidywanie i tak zazwyczaj wystarczy.

Czy jest to idealny model? Nie i trochę mu do niego brakuje ponieważ, jak wspomniane zostało powyżej mamy różne poziomy zaawansowania, różne style uprawiania tego sportu, różne kształty latawców, różne deski, których rozmiar też ma wpływ na generowaną moc itd. Czy ten model jest wystarczający? To już pozostawiam do subiektywnej oceny każdego, jednak moim zdaniem daje satysfakcjonujące wyniki (ograniczając się do założeń odnośnie sprzętu).

Należy również wspomnieć, że dane zebrane w sposób w jaki je zebrałem nie są idealne. Jeśli chcielibyśmy uzyskać naprawde bardzo dobre wyniki takiej regresji należałoby zbierać dane od ludzi schodzących z wody, ponieważ mieliby oni pewność, że dany latawiec był (lub nie był) odpowiedni dla danych warunków wiatrowych, a takie odpowiedzi z domu pewnie są

bliskie prawdy, ale jest to dalej zgadywanie. Jako wieloletni instruktor kitesurfingu dostając pytanie jaki latawiec z mojego zestawu wziąłbym na wodę w danych warunkach wiatrowych w większości wypadków udzieliłbym poprawnej odpowiedzi, ponieważ posiadam latawce co $3m^2$, ale raczej ciężko byłoby mi odpowiedzieć na pytanie jaki dokładnie rozmiar latawca wziąłbym w danych warunkach atmosferycznych jeśli miałbym do wyboru każdy możliwy rozmiar.

Wnioski:

- regresja może skutecznie przewidywać rozmiar latawca,
- odpowiednia transformacja zmiennych znacząco poprawia jakość modelu,
- model przewyższa jakością stosowane przez użytkowników wzory eksperckie.

Model może znaleźć zastosowanie praktyczne – np. w aplikacjach mobilnych lub stronach szkółek kitesurfingowych, jednak aby był on faktycznie przydatny należałoby przeprowadzić dokładniejsze badania i uzyskać lepszej jakości dane.