Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Regresja danych dotyczących zarybiania zbiorników wodnych

Sprawozdanie z laboratorium MSiD

Autor

Ryszard Polkowski

nr albumu: **260376**

kierunek: **Informatyka Stosowana**

7 czerwca 2022

Streszczenie

Praca przedstawia program do wyliczania regresji liczby ryb wykorzystanych do zarybiania zbiorników wodnych. Działa on dzięki danym ze strony `https://data.ny.gov`. Dane zostały oczyszczone ze zbędnych wierszy nie posiadających kompletnych informacji oraz z kolumn posiadających dane zależne od innych kolumn. Następnie zostały z nich wylosowane wiersze do predykcji liczby ryb na podstawie przybliżonej daty, zbiornika wodnego, gatunku i rozmiaru ryb. W programie zostały wykorzystane modele regresji liniowej, SVR z modułu sklearn, oraz customowy model wykorzystujący funkcję curve fit z modułu scipy. Następnie dzięki funkcjom mean_squared_error oraz mean_absolute_percentage_error z modułu sklearn zostały wyliczone błedy kwadratowe i procentowe poszczególnych medeli regresji. Na końcu program wyświetla wykres z oryginalnymi wartościami i wartościami wyliczonymi dzięki poszczególnym modelom regresji.

1 Wstęp – sformułowanie problemu

Autor chce przewidzieć liczbę ryb wykorzystanych do zarybiania zbiorników wodnych. Pozwoli mu to na ocenę liczby ryb użytych do zarybienia w przyszłych latach.

2 Opis danych

Wielkość datasetu 26530 wierszy.

Kolumna "Year" - zmienna całkowitoliczbowa, określa rok zarybienia. Zbiór wartości: 2011-2020.

Kolumna "Waterbody" - zmienna kategoryczna, określa zarybiany zbiornik wodny. Zbiór wartości to 1507 stringów będących nazwami zbiorników wodnych.

Kolumna "Month" - zmienna kategoryczna, określa miesiąc zarybienia. Zbiór wartości: January, February, March, April, May, June, July, August, September, October, November, December.

Kolumna "Number" - zmienna całkowitoliczbowa, określa liczbę ryb wykorzystanych przy zarybieniu. Zbiór wartości: 3-191583000.

Kolumna "Species" - zmienna kategoryczna, określa gatunek zarybianych ryb. Zbiór wartości: Gilt Darter, Lake Herring (Cisco), Steelhead, Rainbow Trout, Chinook, Cisco, Muskellunge, Coho, Sauger, Paddlefish, Tiger Muskellunge, Lake Trout, Lake Sturgeon, Round Whitefish, Brook Trout, Brown Trout, Splake, Walleye, Northern Sunfish, Landlocked Salmon, Panfish.

Kolumna "Size" - zmienna zmiennoprzecinkowa, określa średnią wielkość ryb użytych do zarybienia. Zbiór wartości: 0-30,3.

3 Opis rozwiązania

Dane dotyczące zarybiania zostały pobrane ze strony
`https://data.ny.gov/Recreation/Fish-Stocking-Lists-Actual-Beginning-2011/e52k-ymww`.
Zostały one zapisane w postaci dataframe biblioteki Pandas, zawierających 8 cech określających zarybianie.

Po usunięciu zbędnych kolumn, County oraz Town, zależnych od kolumny Waterbody, program losuje kilkaset wierszy danych do dalszej pracy.

Kożystając z modeli linear_model oraz SVR z biblioteki sklearn, a także własnego modelu regresji na wylosowanych danych, dzięki funkcji predict, program wylicza przewidywane liczności ryb uży-

tych do zarybiania dla podanych podanych lat, miesięcy, zbiorników wodnych, gatunku i średniego rozmiaru ryb.

Następnie dzięki bibliotece matplotlib, prawdziwe i wyliczone dane są nanoszone na wykres i wyświetlane.

4 Rezultaty obliczeń

4.1 Plan badań

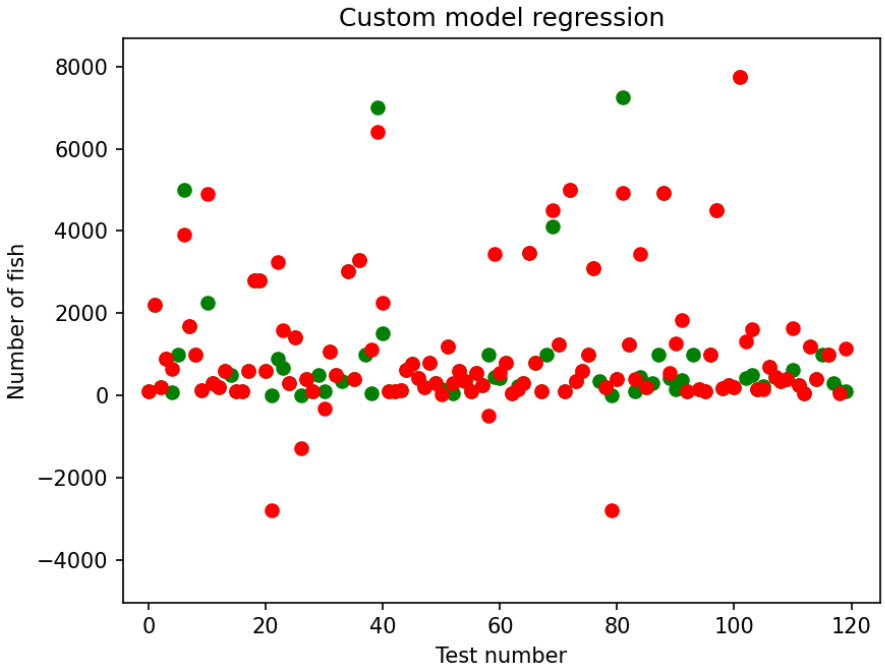
Zbiór danych zostanie podzielony na dwie części: treningową i testową w stosunku 80:20.

4.2 Wyniki obliczeń

Model oceny liczby ryb użytych do zarybiania można przedstawić następującym wzorem:

$$Number = \alpha * Year + \beta * get\_dummies(Waterbody) + \gamma * get\_dummies(Month) + \delta * get\_dummies(Species) + \epsilon * Size + \zeta$$

gdzie get_dummies() to funkcja mapująca dane kategoryczne na reprezentację one-hot. Na rys. 1 pokazany jest przykładowy wykres.



Rysunek 1: Przewidywane wartości dla modelu customowego

Oprócz 4.2 zostały także użyte modele liniowy i SVR

Dzięki funkcjom mean_squared_error oraz mean_absolute_percentage_error z modułu sklearn zostały wyliczone błedy kwadratowe i procentowe poszczególnych medeli regresji, aby sprawdzić ich skuteczność.

5 Wnioski

Przedstawiony program pozwala na dobranie optymalnego modelu regresji do przewidzenia liczby ryb użytych do zarybienia zbiornika wodnego. Po kilku próbach, dla różnych wielkości danych widać skuteczność modelu SVR niezależnie od liości wykorzystanych wierszy, wierszy skuteczność modelu liniowego dla mniejszej liczby wierszy oraz potrzebę większej liczby wierszy dla własnego modelu regresji.

Zależnie od liczby wierszy wziętych do stworzenia modeli, model liniowy osiąga zazwyczaj błąd procentowy(mean absolute percentage error) rzędu kilku procent dla danych nie przekraczających 300 wierszy i kilka milionów procent dla danych przekraczających 300 wierszy.

Model SVR niezależnie od wielkości danych, osiąga błąd procentowy(mean absolute percentage error) od 0,5 do 6 procent. Dla większej liczby danych, model lepiej się uczy i osiąga coraz mniejszy błąd.

Model własny potrzebuje danych posiadających kilkaset wierszy żeby zadziałał, ponieważ liczba kolumn musi być mniejsza niż liczba wierszy, dlatego przy danych posiadających 600 wierszy, model customowy osiąga błąd procentowy(mean absolute percentage error) rzędu kilkudziesięciu procent.

A Dodatek

Kody źródłowe(utrzymane w konwencji języka Python wraz z instrukcjmi uruchomienia) umieszczone zostały w repozytorium github:
`https://github.com/Drysiek/Fish-Stock-Regression`.