Przewidywanie liczby wypożyczeń z miejskiego systemu rowerowego

Projekt zaliczeniowy z przedmiotu „Analiza i przetwarzanie danych w języku Python”
Autor: Filip Baumgart, numer indeksu: 22412,
Informatyka niestacjonarna, III rok, V semestr
Collegium Da Vinci, 2019

1. Opis projektu:

Celem projektu było napisanie programu, który wykorzystując technikę uczenia maszynowego, będzie przewidywał liczbę wypożyczonych rowerów z miejskiego systemu rowerowego. Analizie zostały poddane następujące atrybuty:

- Pora roku,
- Miesiąc,
- Godzina,
- Czy jest to święto,
- Dzień tygodnia,
- Czy jest to dzień roboczy,
- Warunki pogodowe (stopień zachmurzenia, występowanie opadów i/lub mgły, burza),
- Temperatura,
- Temperatura odczuwalna,
- Wilgotność powietrza,
- Prędkość wiatru.

Zbiór danych pochodzi ze strony internetowej:
https://archive.ics.uci.edu/ml/machine-learning-databases/00275/

Repozytorium niniejszego projektu dostępne jest pod adresem:
https://github.com/fbaumgart/Prediction-of-number-of-rentals-from-city-bike-sharing-system

2. Wykonane operacje:

1. Pobranie z serwera pliku .csv zawierającego zbiór danych i wczytanie go jako obiekt typu DataFrame,
2. Przypisanie do utworzonej listy userCount całkowitej liczby wypożyczeń z każdego rekordu zbioru danych,
3. Usunięcie niepotrzebnych kolumn (index, data, rok, liczba zarejestrowanych użytkowników, liczba niezarejestrowanych użytkowników, całkowita liczba użytkowników),
4. Uczenie maszynowe na podstawie przygotowanego zbioru danych przy użyciu algorytmu analizy regresji liniowej,
5. Dopasowanie atrybutów do odpowiednich wyników,
6. Załadowanie zbioru testowego (zbiór testowy to wyizolowane 20% danych z przygotowanego zbioru danych),
7. Przypisanie do utworzonej listy userCounttest całkowitej liczby wypożyczeń z każdego rekordu zbioru danych testowych,

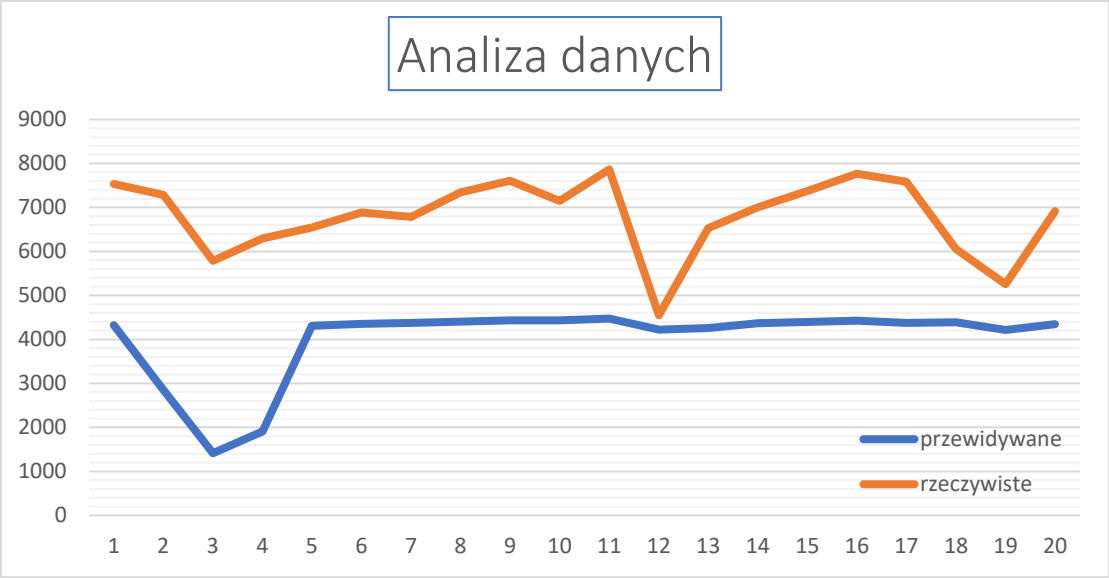
8. Usunięcie niepotrzebnych kolumn (index, data, rok, liczba zarejestrowanych użytkowników, liczba niezarejestrowanych użytkowników, całkowita liczba użytkowników),
9. Przewidywanie liczby wypożyczeń dla rekordu ze zbioru testowego i porównanie go do rzeczywistej wartości ze zbioru testowego

3. Analiza wyników:

Wyniki przewidywania nie są zgodne z rzeczywistymi liczbami, mimo prób zastosowania różnych algorytmów uczenia maszynowego (RidgeRegression, SVR(kernel = „rbf"), SVR(kernel = „linear")). Powodem tej sytuacji jest według mnie zbyt mały zbiór danych (585 rekordów). Wyniki przewidywania mają zbyt niskie wartości co może wskazywać na to, że model jest niedouczony. Współczynnik determinacji r kwadrat uzyskany przez wywołanie metody score() klasy sklearn.svm.SVR ma wartość -0.55.

Przykładowe wyniki:

Przewidziany:	Rzeczywisty:	Różnica:
4324	7534	3210
2851	7286	4435
1413	5786	4373
1910	6299	4389
4312	6544	2232
4350	6883	2533
4376	6784	2408
4402	7347	2945
4432	7605	3173
4435	7148	2713
4473	7865	3392
4220	4549	329
4255	6530	2275
4368	7006	2638
4398	7375	2977
4428	7765	3337
4373	7582	3209
4386	6053	1667
4218	5255	1037
4346	6917	2571


Analiza danych
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
przewidywane
rzeczywiste