

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”), Tree). Powodem tej sytuacji jest według mnie zbyt mały zbiór danych (585 rekordów). Współczynnik determinacji r kwadrat uzyskany przez wywołanie metody `score()` klasy `sklearn.svm.SVR` ma wartość -0.55. Lepsze wyniki w uczeniu maszynowym uzyskano po zastosowaniu klasy `sklearn.tree.DecisionTreeClassifier`. W tym przypadku współczynnik determinacji r kwadrat wynosi 0.00. Mimo tego wyniki nie są zbyt dokładne, co można zobaczyć na wykresie generowanym przez program.