Projekt grupowy

Data Mining

Wykonali:  
Ivan Kaliankovich  
Vladyslav Babych

Spis Treści

[Wstęp 2](#_Toc168421312)

[Zrozumienie uwarunkowań biznesowych 2](#_Toc168421313)

[Zrozumienie danych 2](#_Toc168421314)

[Przygotowanie danych 2](#_Toc168421315)

[Modelowanie 2](#_Toc168421316)

[Ewaluacja 2](#_Toc168421317)

[Wdrożenie 2](#_Toc168421318)

# Wstęp

Systemy wypożyczania rowerów to nowa generacja tradycyjnych wypożyczalni rowerów, w których cały proces od członkostwa, wypożyczenia i zwrotu stał się automatyczny. Dzięki tym systemom użytkownik ma możliwość łatwego wypożyczenia roweru z konkretnego miejsca i oddania go w innym. Obecnie na świecie istnieje ponad 500 programów rowerów publicznych, w których skład wchodzi ponad 500 tysięcy rowerów. Obecnie istnieje duże zainteresowanie tymi systemami ze względu na ich ważną rolę w kwestiach związanych z ruchem drogowym, środowiskiem i zdrowiem.

Oprócz interesujących zastosowań systemów rowerów publicznych w świecie rzeczywistym, charakterystyka danych generowanych przez te systemy czyni je atrakcyjnymi dla badań. W przeciwieństwie do innych usług transportowych, takich jak autobus czy metro, w tych systemach wyraźnie rejestruje się czas podróży, miejsce odjazdu i przyjazdu. Ta funkcja zamienia system rowerów publicznych w wirtualną sieć czujników, którą można wykorzystać do wykrywania mobilności w mieście. Oczekuje się zatem, że monitorowanie tych danych umożliwi wykrycie większości ważnych wydarzeń w mieście.

Do pracy z danymi wykorzystane zostały takie narzędzia jak: Język programowania python, Jupyter notebook, biblioteki sklearn, pandas, seaborn oraz matplotlib.

# Zrozumienie uwarunkowań biznesowych

Naszym celem jest stworzenie modelu regresji przewidującego liczbę wypożyczeń rowerów miejskich dla optymalizacji floty, czyli w oparciu o prognozy zapotrzebowania, firma może dostosować liczbę rowerów dostępnych w różnych lokalizacjach. W przypadku dużego popytu rowery muszą być maksymalnie dostępne dla użytkowników. W przypadku czasu z mniejszym obiegiem rowerów, mogą one być serwisowane.

Przy tworzeniu modelu regresji dla systemu wypożyczalni rowerów, można wziąć pod uwagę różne uwarunkowania biznesowe, które mogą wpływać na liczbę wypożyczeń. Poniżej przedstawiamy kilka przykładowych uwarunkowań biznesowych oraz możliwych zmiennych, które mogą zostać uwzględnione w modelu regresji.

1. Sezonowość:
   * Miesiąc: Liczba wypożyczeń może różnić się w zależności od miesiąca, np. więcej wypożyczeń w cieplejszych miesiącach.
   * Pora roku: Różne pory roku (wiosna, lato, jesień, zima) mogą wpływać na liczbę wypożyczeń.
2. Pogoda:
   * Temperatura: Wyższa temperatura może sprzyjać większej liczbie wypożyczeń.
   * Wilgotność: Wyższa wilgotność może zniechęcać użytkowników do korzystania z rowerów.
   * Prędkość wiatru: Silny wiatr może wpłynąć negatywnie na liczbę wypożyczeń.
   * Warunki pogodowe: Różne warunki pogodowe (np. słonecznie, deszcz, śnieg) mogą wpływać na liczbę wypożyczeń.
3. Dni tygodnia i święta:
   * Dzień tygodnia: Liczba wypożyczeń może różnić się w zależności od dnia tygodnia (np. więcej wypożyczeń w weekendy).
   * Święta: Wypożyczenia mogą wzrosnąć lub zmniejszyć się w zależności od świąt i dni wolnych od pracy.
4. Czas dnia:
   * Godzina: Liczba wypożyczeń może różnić się w zależności od godziny dnia (np. więcej wypożyczeń w godzinach szczytu).

Znaleziona została [baza danych](https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset) uwzględniająca powyższe czynniki.

# Zrozumienie danych

Baza danych posiada co godzinowe pomiary przez dwa lata następujących zmiennych:

    - instant: record index

    - dteday: data

    - season: sezon (1:zima, 2:wiosna, 3:lato, 4:jesień)

    - yr: rok (0: 2011, 1:2012)

    - mnth: miesiąc (1 do 12)

    - hr: godzina (0 do 23)

    - holiday: 1:swięto, 0: zwykły dzień

    - weekday: dzień tygodnia

    - workingday: 1:Ani swięto ani dzień wolny, 0:dzień wolny

    + weathersit:

        - 1: Bezchmurnie, Słabe chmury, Częściowe zachmurzenie, Częściowe zachmurzenie

        - 2: Mgła + Pochmurie, Mgła + Przerywane chmury, Mgła + Niewiele chmur, Mgła

        - 3: Lekki śnieg, lekki deszcz + burza z piorunami + rozproszone chmury, lekki deszcz + rozproszone chmury

        - 4: Ulewny deszcz + palety lodowe + burza + mgła, śnieg + mgła

    - temp: Znormalizowana temperature w stopnia Celsjusza

    - atemp: Znormalizowana oczuwalna temperature w stopnia Celsjusza

- hum: Znormalizowana wilgotność

    - windspeed: Znormalizowana prędkość wiatru

    - casual: ilość aktywnych użytkowników

    - registered: ilość zarejestrowanycj użytkowników

    - cnt: ilość rowerów wypożyczonych

Wypisanie statystyk dla danych  
Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Oprogramowanie multimedialne

Opis wygenerowany automatycznie  
Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Czcionka

Opis wygenerowany automatycznie

Baza danych zawiera 17379 obserwacji, czyli jest zgodna z wymaganiami (>10000 obserwacji).  
  
Po sprawdzeniu brakujących wartości dla obserwacji okazało się, że każdy z pomiarów w ciągu 2 lat został prawidłowo przeprowadzony, gdyż baza danych nie ma brakujących wartości.

W celu zrozumienia danych postanowiliśmy przygotować macierz korelacji  
Obraz zawierający tekst, zrzut ekranu, Równolegle, Prostokąt

Opis wygenerowany automatycznie  
Na powyższym wykresie rzuca się w oczy silna korelacja pomiędzy zmiennymi temp oraz atemp. Postanowiliśmy nie uwzględniać zmiennej atemp w naszym modelu.

Użyliśmy wykresów słupkowych do rozkładu zmiennych kategorycznych, przykład poniżej  
Obraz zawierający tekst, zrzut ekranu, wyświetlacz, Prostokąt

Opis wygenerowany automatycznie

Obraz zawierający zrzut ekranu, Prostokąt, diagram, design

Opis wygenerowany automatycznie  
W trakcie świąt rzadko kiedy zdarza się duża liczba rowerów wypożyczonych.

# Przygotowanie danych

Usunęliśmy z modelu zmienne wejściowe dteday, ponieważ bardziej od daty, potrzebowaliśmy informacji dotyczące dnia tygodnia oraz godziny.  
Po analizie macierzy korelacji postanowiliśmy usunąć również zmienne mnth oraz atemp.  
Obraz zawierający tekst, Czcionka, zrzut ekranu, Grafika

Opis wygenerowany automatycznie

Atemp jest przesadnie skorelowana ze zmienną temp i nie potrzebujemy obu. Zmienna wejściowa mnth również posiada mocną korelację ze zmienną season, ale season bardziej wpływa na ilość rowerów wypożyczonych i jest dużo mniej złożona (4 możliwych stanów zamiast 12).

Mieliśmy dużo dyskusji w zespole jak postąpić ze zmiennymi kategorycznymi np. taką jak godzina pomiaru, ale zdecydowaliśmy się na one-hot encoding za pomocą metody get\_dummies biblioteki pandas.

Twórcy bazy danych znormalizowali zmienne wejściowe ciągłe metodą MinMax, czyli każda ze zmiennych w naszym modelu przed budowaniem regresji jest w zakresie od 0 do 1.

# Modelowanie

# Ewaluacja

# Wdrożenie