Projekt grupowy

Data Mining

Wykonali:  
Ivan Kaliankovich  
Vladyslav Babych

Spis Treści

[Wstęp 2](#_Toc168604178)

[Zrozumienie uwarunkowań biznesowych 2](#_Toc168604179)

[Zrozumienie danych 3](#_Toc168604180)

[Przygotowanie danych 6](#_Toc168604181)

[Modelowanie 7](#_Toc168604182)

[Ewaluacja 7](#_Toc168604183)

[Modelowanie 2 8](#_Toc168604184)

[Ewaluacja 2 9](#_Toc168604185)

[Wdrożenie 10](#_Toc168604186)

# 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

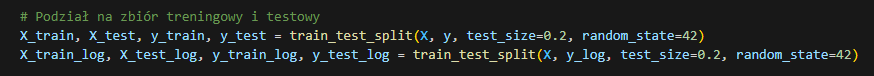
Usunęliśmy z modelu zmienne wejściowe dteday, ponieważ bardziej od daty, potrzebowaliśmy informacji dotyczące dnia tygodnia oraz godziny.  
Po analizie macierzy korelacji postanowiliśmy usunąć również zmienne mnth oraz atemp.  
Obraz zawierający tekst, Czcionka, zrzut ekranu, Grafika

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.

Po obróbce danych podzieliliśmy dane na zbiory treningowe i testowy w stosunku 8 do 2.  


# Modelowanie

Za pomocą biblioteki sklearn udało się na podstawie przetworzonych danych stworzyć model regresji logarytmicznej.   
Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

Poniżej przedstawiamy wykres rozproszenia pierwszego modelu.

Obraz zawierający tekst, zrzut ekranu, Wykres, linia

Opis wygenerowany automatycznie

# Ewaluacja

Poniżej przedstawiamy podsumowanie modelu regresji liniowej z najważniejszymi metrykami omawianymi na wykładzie  
Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Czcionka

Opis wygenerowany automatycznie  
R-squared (R²) i Adjusted R-squared są identyczne na poziomie: 0.682. Oznacza to, że model wyjaśnia 68,2% zmienności zmiennej zależnej (cnt). Uzyskany wynik uznaliśmy za stosunkowo dobry jak na pierwszy model. Adjusted R² jest równy R², co oznacza, że dodanie dodatkowych zmiennych nie wprowadza nadmiernego dopasowania.  
Inne statystyki mogą sugerować, że reszty modelu mogą nie być idealnie normalnie rozłożone, co może wpływać na niektóre założenia modelu.  
Posiadamy również P-value wszystkich zmiennych wejściowych modelu, niestety nie mieszczą się na zrzucie ekranu, więc dołączymy je jako osobny plik w wersji tekstowej.  
Większość zmiennych ma p-values poniżej 0.05, co oznacza, że są istotne statystycznie. Kilka zmiennych, takich jak holiday, summer, weekday\_1, weekday\_2, weekday\_3, weekday\_4, Light Snow, Heavy Rain, mają wyższe p-values, sugerując, że mogą nie być istotne w modelu.

Po głębszej analizie wyników postanowiliśmy spróbować logarytmowania zmiennej wyjściowej (cnt) i sprawdzenie również zlogarytmowanego zbioru testowego na nowo nauczonym modelu regresji liniowej.

# Modelowanie 2

Stworzyliśmy nowy zestaw danych logarytmując zmienną wyjściową.  
Obraz zawierający tekst, zrzut ekranu, oprogramowanie, Czcionka

Opis wygenerowany automatycznie

Identycznie nauczyliśmy model do przypadku poprzedniego. Uzyskaliśmy następujący wykres rozproszenia:

Obraz zawierający tekst, zrzut ekranu, linia, Wykres

Opis wygenerowany automatycznie  
Wykres na pierwszy rzut oka był bardziej zgodny z oczekiwaniami, szczególnie dla przypadków wypożyczenia:

# Ewaluacja 2

Następnym krokiem jest porównanie modelu do poprzedniego. Wyniki w znaczącym stopniu się poprawiły.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

R²: 0.824  
Adjusted R²: 0.823

Wartości R² i Adjusted R² są wyższe w porównaniu do poprzedniego modelu (wcześniej 0.682). To sugeruje, że nowy model wyjaśnia większą część zmienności zmiennej zależnej, co wskazuje na lepsze dopasowanie. Wartość F-statistic jest wyższa (1621 > 744.8), co również sugeruje lepsze dopasowanie modelu. Wartości współczynników Log-Likelihood, AIC oraz BIC wszystkie wskazują na lepsze dopasowanie modelu.  
Wartości współczynników wskazują na istotne zmiany w niektórych zmiennych. Większość zmiennych jest istotna statystycznie (p-value < 0.05).

* Zmienna **temp** ma nadal znaczący pozytywny wpływ na liczbę wypożyczeń (1.5644).
* **hum** i **windspeed** mają negatywny wpływ.
* **holiday** jest teraz istotna statystycznie i ma pozytywny wpływ

Statystyki Omnibus, JB oraz Skew i Kurtosis sugerują, że dane mogą mieć pewne odstępstwa od normalności, ale nie są one nadmierne.

Podejmowaliśmy dalsze próby polepszenia modelu takie jak: progowanie zmiennych wejściowych, usunięcie jedynej zmiennej o większej P-value od 0.05 (Heavy Rain), ale żadna z metod nie polepszyła znacząco modelu. Rozważenie o usunięcie wpływowych obserwacji odstających jest kolejnym dobrym pomysłem, natomiast takie podejście wymaga dużej ilości poświęconego czasu.

# Wdrożenie

Poniżej przedstawiamy kolejne kroki w stronę wdrożenia modelu w rzeczywiście działający biznes.

* Przywrócenie wartości przewidywanych wypożyczeni do rzeczywistej wartości (przed logarytmowaniem)
* Integracja modelu z systemem zarządzania wypożyczalnią rowerów.
* Zawarcie danych o użytkowniku w poważnym systemie chmurowym
* Dodanie backupu dla danych jak i modelu
* Aktualizacja modelu dalej i próba uzyskania jak najlepszych wyników prognoz.

Korzyści z użycia zaproponowanego przez nas rozwiązania**:**

* Optymalizacja zasobów: Model może pomóc w uniknięciu niedoborów rowerów w szczytowych momentach i nadmiaru rowerów w okresach słabego ruchu.
* Poprawa obsługi klienta: Klienci będą bardziej usatysfakcjonowani, jeśli będą mogli wypożyczyć rower bezproblemowo, bez długich kolejek oczekiwania.
* Zwiększenie zysków: Dokładne prognozy zapotrzebowania mogą pomóc w optymalizacji cen i zwiększeniu przychodów.
* Lepsze planowanie: Model może pomóc w planowaniu działań marketingowych, rekrutacji pracowników i konserwacji rowerów.
* Wzrost efektywności: Model może pomóc w identyfikacji trendów i czynników wpływających na zapotrzebowanie na rowery, co może prowadzić do lepszego zarządzania firmą.