Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Predykcja ilości wypożyczeń roweru miejskiego we Wrocławiu

Sprawozdanie z laboratorium

Autor

Michał Sternik

nr albumu: **260455**

kierunek: **Informatyka Stosowana**

14 czerwiec 2022

Streszczenie

Dokument dotyczy aplikacji predykującej tygodniową średnią kroczącą ilości wypożyczeń rowerów miejskich we Wrocławiu. Dane zostały pobrane z internetowego serwisu https://www.wroclaw.pl/open-data/, jednak ze względu na liczne błędy związane z pomiarami, by odpowiednio przygotować zbiór danych należało połączyć dwa odrębne datasety oraz odpowiednio je przekonwertować do postaci takiej, by było możliwe ich złączenie. Pierwszy dataset jest pobierany i aktualizowany na bieżąco (codziennie) używając pakietu beautifulsoup4, a kolejny został pobrany jednorazowo i stanowi trzon aplikacji. Podjęte zostały próby kontaktowania się telefonicznie oraz mailowo z działem technicznym odpowiedzialnym za udostępniane datasety, jednak bezskutecznie, zatem nie udało się uprościć implementacji do skorzystania z jednego datasetu. Opis danych oraz modeli szczegółowo opisany zostanie kolejno w podpunkcie 2 oraz 3.

1 Wstęp – sformułowanie problemu

Jaki jest problem? Autor interesuje się zagadnieniem transportu publicznego we Wrocławiu i pragnie sprawdzić jak duży udział na liczbę wypożyczeń rowerów mają takie parametry jak aktualna pogoda oraz sezonowy klimat panujący w Polsce.

2 Opis danych

Dane są pobierane w plikach opisujących każde pojedyncze wypożyczenie na przestrzeni jednego dnia, bądź całego miesiąca. Najwcześniejszy rekord datasetu datowany jest na 1.01.2020, a ostatni na 14.06.2022. W każdy dzień do datasetu doliczanych jest średnio 5 tysięcy rekordów. Wielkość pierwotnego datasetu zatem to około 4 miliony rekordów. Nas jednak nie interesuje każdy rekord z kolei, a łączna ich suma w każdym dniu co w ostatecznym rozrachunku daje nam 838 rekordów. Każdy rekord opisany jest przez: dzienną liczbę wypożyczeń (typ Int), średnią temperaturę w danym dniu (float), wielkość opadu atmosferycznego w danym dniu (w milimetrach na metr kwadratowy, float), wartość wypożyczeń z dnia poprzedniego (int), oraz kodowania one hot dla miesięcy co daje dodatkowe 12 kolumn (styczen-listopad, typ w kolumnie: int). Ostateczne wymiary datasetu to 838x17 (stan na 13.06.2022)*

*z datasetu zostały wyłączone dane z lockdownu covidowego - 25.03.2020 - 10.04.2020

3 Opis rozwiązania

Dane zostały pobrane ze strony `https://www.wroclaw.pl/open-data/`. Dostęp do danych został zrealizowany pakietem beautifulsoup4 do pobierania danych na bieżąco (od 2022 roku codziennie), a dane z lat 2020, 2021 zostały pobrane z internetu ręcznie. Oba pobrane datasety zostały połączone w jeden biblioteką `Pandas` i łączą się razem w kompletny dataset od początku 2020 roku do dziś. Używając metody regresji liniowej na danych opisanych w podpunkcie 2, uzyskano model pozwalający dopasować się do danych wejściowych i zauważyć widoczną sezonowość w ilości przejazdów. W dni letnie gdzie średnia temperatura jest wyższa ilość przejazdów była znacząco wyższa niż w dni zimowe z niższą temperaturą.

4 Rezultaty obliczeń

4.1 Plan badań

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

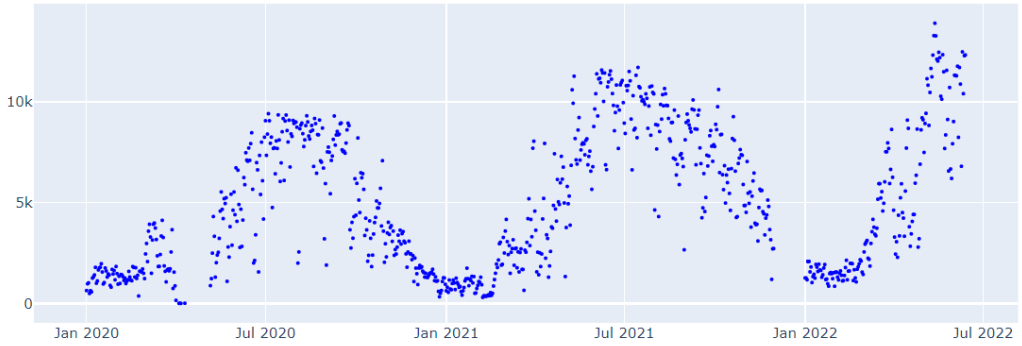
4.2 Wyniki obliczeń

Model predykcji ilości przejazdów można przedstawić następującym wzorem:

$$iloscPrzejazdow7dAvg = a*x[”tavg”]+b+x[”prcp”]+d*x[”prevValue”]+oneHotEncoding(months)+q$$

(1)gdzie $tavg$ to średnia temperatura dzienna, $prcp$ to dzienna suma opadów w milimetrach, $prevValue$ to wartość z dnia poprzedniego.

Na rys. 1 pokazany jest wykres z danymi od początku 2020 do czerwca 2022. Po zaaplikowaniu



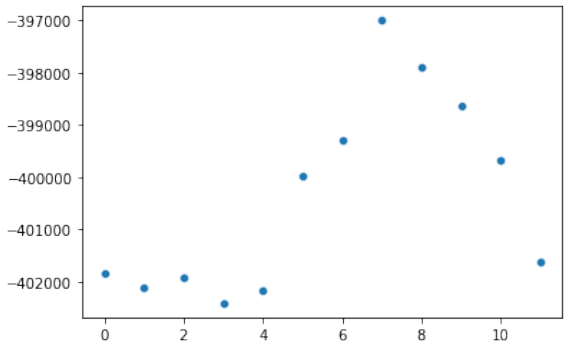
Rysunek 1: Dzienne przejazdy roweru miejskiego (liczba przejazdów od daty)

Po zaaplikowaniu metody curve.fit do naszych danych testowych otrzymujemy następujące wyniki (rys .2):

0	-0.097212
1	23.874712
2	0.327187
3	-401822.849422
4	-402121.301993
5	-401908.087052
6	-402407.454429
7	-402167.050168
8	-399973.697965
9	-399294.347294
10	-397004.288513
11	-397904.786544
12	-398645.429180
13	-399682.363439
14	-401627.630941
15	403220.945362

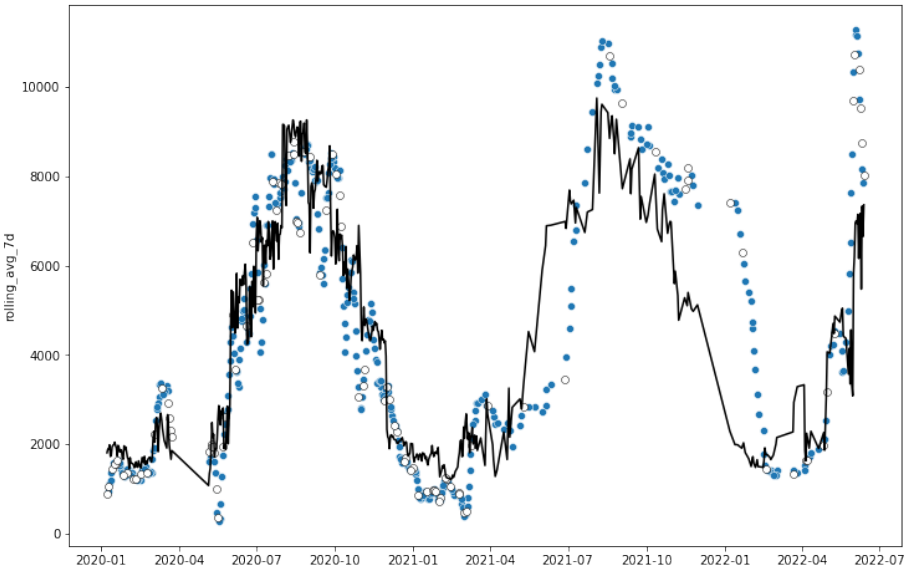
Rysunek 2: Rysunek 2: Parametry po wywołaniu funkcji $curve_fit$

Widzimy, że parametr o indeksie 0 (tavg) oraz 2 (prevValue) nie wnoszą dużo do modelu. Parametr o indeksie 1 (prcp) wnosi nieco więcej od pozostałych, najwięcej wnosi za to kodowanie oneHot miesięcy, dzięki którym zauważamy pewną sezonowość w modelu. Wykres kodowania (Rysunek 3) one hot miesięcy, na którym zauważa się sezonowość prezentuje się następująco:



Rysunek 3: Rysunek 3 Parametry dla kodowania oneHot (widać sezonowość dla miesięcy letnich)

Końcowy wykres z nałożonym własnym modelem regresji prezentuje się następująco:



Rysunek 4: Model regresji nałożony na gotowy wykres 7-dniowej średniej kroczącej wypożyczeń od daty)

Po obliczeniu błędu średniokwadratowego, który wyniósł (dla aktualnie sprawdzanych danych testowych) `2739152.9211411052`, możemy zauważyć, że model myli się średnio o 1655 rekordów dziennie.

5 Wnioski

Przedstawiona metoda pozwala na otrzymanie modelu, który jest wrażliwy na nietypowe dane, na przykład dla 1 września gdy sezonowo trend mniejszej ilości wypożyczeń się powtarzał, błąd średniokwadratowy mocno rósł. Podobnie z innymi charakterystycznymi dniami w roku. Po usunięciu danych z okresu covidowego błąd dodatkowo wzrósł, mimo że zostało usunięte zaledwie kilkanaście rekordów. Mimo to model spełnia swoją rolę - wykazuje sezonowość danych - widać rekordowe ilości wypożyczeń dla lipca i sierpnia, a minimalne dla okresu zimowego, od grudnia do marca.

6 Dodatek

Kody źródłowe z wykorzystaniem bibliotek obrazowania oraz przetwarzania danych (pandas, numpy, seaborn, datetime, os, requests, beautifulsoup4, plotly) zostały umieszczone w repozytorium github: `https://github.com/michal-sternik/wroclaw-bike-share-analysis`.