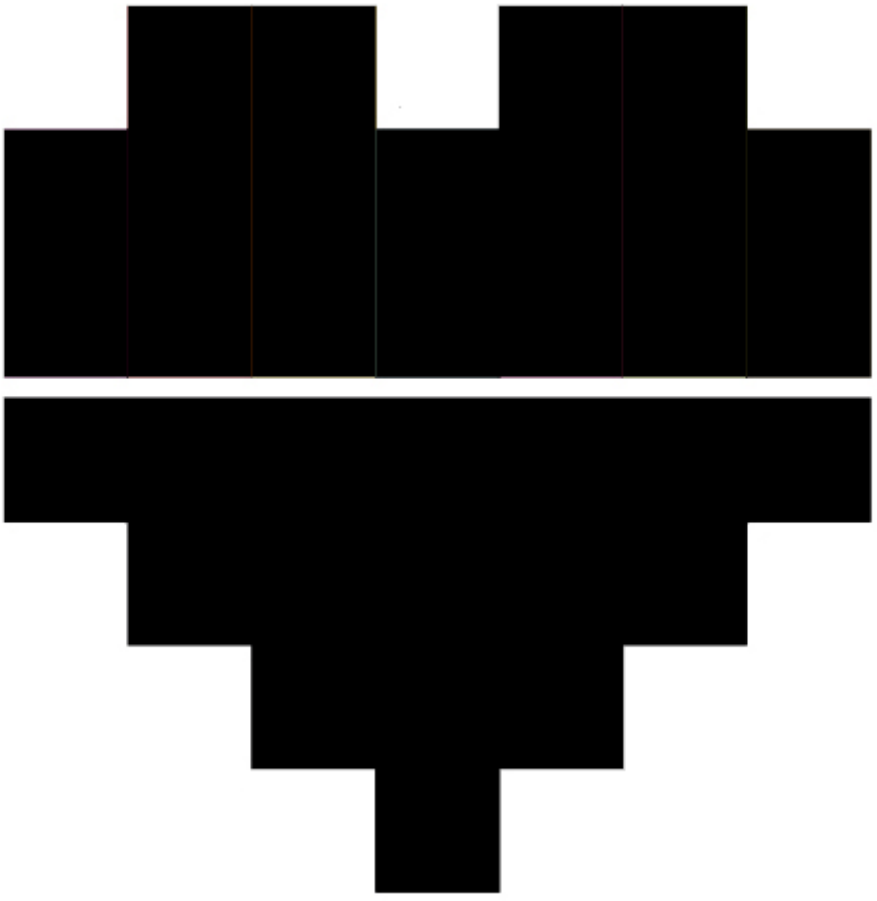
DATAWORKSHOP

KATOWICE TEAM



Cel projektu

Celem projektu było zbadanie, w jaki sposób mieszkańcy miasta oraz osoby przyjezdne użytkują rowery miejskie - w jakie dni pojazdy najczęściej są wypożyczane, dokąd najchętniej jeżdżą użytkownicy, ile czasu przeznaczają na jazdę na rowerze miejskim, czy rowery są wykorzystywane do rekreacji, czy wyłącznie do transportu. Podjęto też próbę prognozowania liczby wypożyczeń rowerów za pomocą algorytmów uczenia maszynowego.

Wprowadzenie

O City by bike

System rowerów miejskich City by bike w Katowicach zadebiutował w 2015 roku. Początkowo mieszkańcy korzystali z trzech stacji, rok później 11, a w 2017 roku było ich 35. Do wypożyczenia roweru wystarczy telefon komórkowy i rejestracja w aplikacji, a wszystkie przejazdy trwające mniej, niż 15 minut są darmowe. Utworzenie systemu wypożyczalni rowerów miejskich zostało powierzone Przedsiębiorstwu Komunikacji Miejskiej Katowice Sp. z o.o, a całościowy koszt projektu zawierający dzierżawę urządzeń, rowerów oraz serwis wyniósł 1,2 mln zł. Poza rozwojem sieci rowerów miejskich w Katowicach duże środki są inwestowane w rozbudowę infrastruktury rowerowej - na przestrzeni ostatnich czterech lat w Katowicach powstało 15,5 km nowych dróg rowerowych.

Pozyskane dane

Dane na temat wykorzystania rowerów miejskich zostały pobrane z serwisu otwartedane.medialabkatowice.eu. Operator systemu, firma Nextbike, przekazał dane w arkuszu kalkulacyjnym, z którego przed publikacją usunięto niepełne rekordy. Pliki zawierające dane dotyczące przejazdów rowerowych zawierały następujące kolumny:

- bike_num - numer identyfikacyjny roweru
- start_time - godzina wypożyczenia
- end_time - godzina zwrotu
- departure - stacja z której rower został wypożyczony
- return - stacja do której rower został zwrócony
- duration_sec - czas wypożyczenia

Natomiast pliki zawierające dane dotyczące stacji rowerowych kolumny:

- name - nazwa
- id - numer identyfikacyjny stacji
- lat - szerokość geograficzna
- lon - długość geograficzna
- capacity - pojemność (liczba stojaków na rowery) stacji

Współrzędne geograficzne stacji pobrano z serwisu OpenStreetMap.org oraz innych źródeł.
Dane o **pyle zawieszonym PM10 i PM 2.5** (smog) ze strony http://powietrze.gios.gov.pl/pjp/archives
Dane te były uśrednione za dany dzień i wyrażone wartością ciągłą z przedziału 0-112 (dla PM10) oraz 0-88 (PM 2.5)

Dane o **pyleniu topól, brzóz, traw** (roślin występujących w miastach). Jako że przez większość roku coś pyli, brano pod uwagę tylko wysokie stężenia pyłków w/w roślin, przyjmując jako 1 że dana roślina pyli (w wysokim stężeniu), lub 0 (nie pyli w wysokim stężeniu). Dane pozyskane z https://www.claritine.pl/pl/prognoza-dla-alergikow/kalendarz-pylenia-sprawdz-aktualne-pylenie/

Dane na temat dni, które odbywał się **mecz reprezentacji Polski** w piłce nożnej pozyskane ze strony http://www.hppn.pl/reprezentacja/mecze/rok-2018?c=1 - chcieliśmy sprawdzić, czy mieszkańcy miast mogą rowerami docierać do ogródków piwnych i miejsc z telebimami.

Dane pogodowe z publicznego zbioru danych dostępnego w Google BigQuery udostępnionego przez National Oceanic and Atmospheric Administration (NOAA) - https://console.cloud.googlecom/marketplace/details/noaa-public/gsod.

Analiza danych

Czyszczenie danych

Wypożyczenia powyżej 12 godzin

Ponieważ operator dopuszcza wypożyczenie roweru trwające do 12 godzin, wszystkie wypożyczenia trwające dłużej zostały potraktowane jak błąd i wyrzucone ze zbioru danych. Było ich zaledwie 78 i stanowiły one 00.045% zbioru.

Wypożyczenia poniżej 3 minut

W pozyskanych danych brak jest informacji na temat uszkodzeń rowerów lub powodów zwrotu. Dlatego nasz zespół założył, że jeżeli wypożyczenie było krótsze niż 3 minuty, a stacja wypożyczenia była taka sama, jak stacja zwrotu, to użytkownik rozmyślił się, bądź rower był wadliwy i został oddany. Wszystkie takie rekordy (10.5% całości) zostały usunięte z bazy. Następnie zbadano wszystkie wypożyczenia, które były krótsze niż 3 minuty (1777 rekordów), ale stacja wypożyczenia była różna od stacji oddania roweru. Według strony Google Maps w jedną minutę rower może przejechać około 400 metrów, zmierzono więc odległość między stacjami dla każdego wypożyczenia i jeśli była ona większa, niż liczba minut pomnożona przez 400 metrów, rekordy zostały usunięte (186 wierszy).
W sumie usunięto 20549 wierszy, czyli 10.6% całego zbioru, do analizy pozostało natomiast 172701 rekordów

Analiza

Dane ogólne

Całkowita liczba rowerów, które poruszały się przez okres działania stacji po mieście: **420**
Przez **260** dni zarejestrowano **172701** wypożyczeń, co daje średnio **664** wypożyczeń na dzień.
Średni czas wypożyczenia: **33,22** minuty
Liczba wypożyczeń, które trwały poniżej 15 minut (a więc były darmowe): **93823** - **54%**
Liczba wypożyczeń, które trwały:

do 1h:	**14534**
między 1h a 2h	17859
między 2h a 3h	6002
między 3h a 4h	2196
między 4h a 5h	741
między 5h a 6h	254
między 6h a 7h	117
między 7h a 8h	59
między 8h a 9h	32
między 9h a 10h	29
między 10h a 11h	22
między 11h a 12h	22

Suma wypożyczeń w poszczególnych miesiącach:

4	**24719**
5	26299
6	28518
7	26896
8	29413
9	17326
10	12140
11	6227
12	1163

Dni tygodnia

Suma wypożyczeń w poszczególne dni tygodnia:

poniedziałek	**24813**
wtorek	24504
środa	25667
czwartek	24564
piątek	24069
sobota	23736
niedziela	25348

Średni czas wypożyczenia (w minutach) w poszczególne dni tygodnia:

poniedziałek	**27.27**
wtorek	27.02
środa	27.62
czwartek	29.68
piątek	27.32
sobota	41.91
niedziela	51.6

Dni robocze vs weekendy i święta:

średnia liczba wypożyczeń w dni robocze: **668.2**

średnia liczba wypożyczeń w weekendy: **654.45**

średnia liczba wypożyczeń w dni świąteczne: **674.7**

średni czas wypożyczeń w dni robocze: **27.78**

średni czas wypożyczeń w weekendy**: 46.91**

średni czas wypożyczeń w dni świąteczne: **56.57**

W czasie dni roboczych liczba wypożyczeń rowerów rośnie między godziną 6 i 8 rano, a następnie między 14 a 17, co pozwala przypuszczać, że część użytkowników to osoby dojeżdżające rowerami do miejsc pracy. W ciągu weekendów i świąt natomiast najpopularniejsze pory wypożyczeń to popołudnia.

Dni w których odnotowano najwięcej wypożyczeń:

2018-08-15	**1331**
2018-06-20	1323
2018-06-06	1309
2018-05-31	1291
2018-06-10	1269

Najpopularniejsze trasy 15 sierpnia 2018 (najpopularniejszy dzień):

Dolina 3-ch Stawów - Dolina 3-ch Stawów

Katowice Rynek - Katowice Rynek

Al. Bolesława Krzywoustego - Al. Bolesława Krzywoustego

Al. Księżnej Jadwigi Śląskiej - Al. Księżnej Jadwigi Śląskiej

KTBS – Saint Etienne 1 - KTBS – Saint Etienne 1

W których dniach popularnych miesięcy (od kwietnia do września) było najmniej wypożyczeń:

2018-05-17	**105**
2018-07-18	245
2018-07-17	301
2018-06-28	380
2018-05-18	387

Niedziele handlowe vs niehandlowe

liczba niedziel handlowych w okresie działania roweru miejskiego: **9**
liczba niedziel niehandlowych w okresie działania roweru miejskiego: **29**
wypożyczenia w niedziele handlowe: **3879**
wypożyczenia w niedziele niehandlowe: **21469**
średnia l. wypożyczeń w niedziele handlowe: **431.0**
średnia l. wypożyczeń w niedziele niehandlowe: **740.31**
średni czas wypożyczenia w niedziele handlowe: **44.44**
średni czas wypożyczenia w niedziele niehandlowe: **52.89**

Stacje rowerowe

Liczba stacji w Katowicach: **54**

Najpopularniejsze stacje wypożyczenia:

Katowice Rynek
Silesia City Center
KTBS – Krasińskiego
Murapol Mariacka
Al. Bolesława Krzywoustego

Najpopularniejsze stacje zwrotu:

Katowice Rynek
KTBS – Krasińskiego 14
Murapol Mariacka
Silesia City Center
Al. Bolesława Krzywoustego

Najmniej popularne stacje wypożyczenia:

PKN Orlen - Al. Roździeńskiego
ING Roździeńska
PKN Orlen - Bocheńskiego
PKN Orlen - Piotrowicka
PKN Orlen - Murckowska

Najmniej popularne stacje zwrotu:

ING Roździeńska
PKN Orlen - Al. Roździeńskiego
PKN Orlen - Bocheńskiego
PKN Orlen - Murckowska

Najpopularniejsze trasy:

Katowice Rynek - Katowice Rynek
KTBS – Krasińskiego 14 - Katowice Rynek
Dolina 3-ch Stawów - Dolina 3-ch Stawów
Katowice Rynek - KTBS – Krasińskiego 14
Al. Księcia Henryka Pobożnego - Al. Księcia Henryka Pobożnego

Najpopularniejsze trasy ze stacją zwrotu różną od stacji wypożyczenia:

KTBS – Krasińskiego 14 - Katowice Rynek
Katowice Rynek - KTBS – Krasińskiego 14
Silesia City Center - Katowice Rynek
Katowice Rynek - Murapol Mariacka
Katowice Rynek - Silesia City Center

Rowery

Najbardziej eksploatowany w czasie sezonu 2018 rower został wypożyczony **774** razy i odwiedził 42 z 54 stacji. Biorąc pod uwagę liczbę wszystkich rowerów i wszystkich wypożyczeń, każdy z rowerów został wypożyczony średnio 3 razy.

Analiza predyktywna

Dysponując wcześniej przygotowanymi danymi postanowiliśmy sobie za cel przeprowadzić prognozę liczby wypożyczeń dla danego dnia sumarycznie w całym mieście. Hipotetyczny cel prognozy to wspomożenie administratorów stacji w przewidywaniu ile rowerów mogą maksymalnie zdjąć ze stacji w celach serwisowych danego dnia.

Dla tak postawionego zadania, zbiór danych został odpowiednio ograniczony, tak aby zawierał kolumny które wydają się mieć jakiś wpływ na sumę wypożyczeń w danym dniu dla całego miasta. Dodatkowo wzbogaciliśmy dane informacją o pogodzie ze wspomnianego wcześniej datasetu. Dokładną listę kolumn można znaleźć w udostępnionym notebooku Jupytera.

Przed przystąpieniem do analizy za pomocą algorytmów uczenia maszynowego, wyznaczyliśmy tkzw benchmark na podstawie średniej liczby wypożyczeń (672) i obliczonego dla niej błędu RMSE który wyniósł 350.

Kolejny krok to podział zbioru danych na część testową i treningową. Ze względu na niewielką ilość danych (254 wiersze) podział nie był zupełnie przypadkowy i wykorzystaliśmy technikę zwaną z języka angielskiego stratified sampling. Dane podzielone są na jednorodne grupy zwane warstwami, celem jest to, aby z każdej warstwy pobrana została odpowiednia liczba przypadków, aby nasz zbiór danych był dobrą reprezentacją danych ogólnych. Do stworzenia grup wybraliśmy informację o opadach atmosferycznych, jako dość istotną a jednocześnie dla zdecydowanej większości przypadków mówiącą o braku opadów. Docelowo, chcieliśmy aby zarówno wzbiorze treningowym jak i testowym znalazły się obserwacje z pełnym rozkładem opadów atmosferycznych.

W kolejnym kroku dane zostały przygotowane pod analizę algorytmami uczenia maszynowego, w tym celu:

- dane kategoryczne zostały przekształcone za pomocą “One Hot Encodera”,
- dla kolumn liczbowych zapewniliśmy się, że brakujące dane są uzupełnione średnią,
- dane liczbowe zostały także ustandaryzowane.

Transformacje zostały przeprowadzone niezależnie dla zbioru treningowego a następnie testowego, mając na uwadzę aby dopasowanie zostało przeprowadzone tylko na zbiorze treningowym.

W kolejnym kroku sprawdziliśmy kilka modeli uczenia maszynowego, obliczyliśmy RMSE, a następnie do liczenia RMSE wykorzystaliśmy kross walidację.

Wykorzystaliśmy następujące algorytmy:

- regresja liniowa,
- drzewo decyzyjne,
- las losowy,
- XGBoost,
- Support Vector Regression.

Aby wyniki poszczególnych modeli bardziej odpowiadały rzeczywistości do obliczania błędu (RMSE) użyliśmy kroswalidacji.

Ostatnim etapem było tuningowanie najbardziej obiecujących modeli za pomocą funkcji GridSearchCV.

W efekcie udało się nam uzyskać następujące wyniki (RMSE):

- regresja liniowa: 174
- las losowy: 150
- XGBoost: 155
- Support Vector Regression: 300

Najlepszym modelem okazał się las losowy który pozwolił na obniżenie błędu względem "benchmarku" o 200.

Do poprawy:

- Można spróbować ustandaryzować i znormalizować dane. To sprawi, że model nie będzie widział pewnych cech jako ważniejszych (bo np. 90 (stopni Fahrenheita) > 6 (dayofweek – dzień tygodnia) > 0.18 (rain – opady w litrach / m2). Można wykorzystać np. StandardScaler z biblioteki preprocessing.
- Można spróbować usunąć cechy nieznaczące i skorelowane ze sobą. Np. jak cecha visib (widoczność) jest zależna od fog (mgła).

- Dla każdej cechy można utworzyć histogram aby stwierdzić, czy rozkład posiada długi ogon. Jeśli takowy występuje, to można wykorzystać np. 95 percentyl do ograniczenia zakresu danych.
- Alternatywnie można spróbować zlogarytmować daną cechę i dodać 1 przed użyciem w modelu. Operacja odwrotna to funkcja wykładnicza -1. Taką operację warto stosować gdy jest duży rozrzut wartości.
- Zapełnić luki gdzie nie zarejestrowano wartości pomiaru dla danego dnia. Można w tym celu wykorzystać metodę fillna dla datasetu.
- Można spróbować wszystkie cechy hot-encodować. Założenie jest takie, że modele lepiej radzą sobie z danymi binarymi, niż z wartościami ciągłymi z szerokiego przedziału.

Czego się nauczyliśmy

Poprzez udział w projekcie nasza grupa sympatyków Machine Learning poznała dogłębnie możliwości pakietów analitycznych Pandas, Numpy. Projekt Rowery pochłaniał nas na spotkaniach i poza nimi. Godzinami wpatrywaliśmy się w monitory próbując wyciągnąć jak najwięcej informacji z danych. Zmagania rozpoczęliśmy od solidnej eksploracji zestawu danych jaki został nam udostępnionych przez firmę nextBike. Pierwsze kroki jakie chcieliśmy wykonać to zrozumienie jak, gdzie, kiedy przemieszczają się mieszkańcy Katowic korzystający z notabene nowego środka komunikacji ja kim są rowery miejskie. Analizowaliśmy najczęściej wybierane trasy, czas podróży, ilość przesiadek, przystanki znacznie oddalone od innych oraz wiele innych kwestii związanych z wycieczkami rowerowymi. To wszystko doprowadziło nas do etapu kiedy stwierdzimy, że nasza wiedza jest wystarczająco duża, aby usiąść do prognozowania ilości wypożyczeń na przyszłym sezon, co od samego początku było naszym głównym celem w tym projekcie. Nim przystąpiliśmy do sedna, nauczyliśmy się kolejnej ważnej rzeczy jaką jest feature engineering. Przekształcaliśmy istniejące zmienne, tworzyliśmy nowe dane, a wszystko po to, żeby nasz model mógł dostarczyć jeszcze lepsze wyniki. Jak się później okazało pobieranie danych z źródeł zewnętrznych, takich jak temperatura powietrza, informacja o imprezach masowych w okolicy (mecz piłkarski, koncert) miało bardzo pozytywny wpływ na nasz model. Gdy już mieliśmy przygotowany cały zestaw jaki będziemy chcieli użyć rozpoczęliśmy prace związane z prognozowaniem. Na tym etapie poznaliśmy klasyfikatory takie jak regresja liniowa, drzewa decyzyjne, lasy losowe czy sieci neuronowe. Budowaliśmy wiele modeli, a nawet zorganizowaliśmy konkurs na najlepszy model, dzięki czemu poznaliśmy Kaggle, a zwycięzca w ramach nagrody dostał wejściówkę na konferencję organizowaną przez DataWorkshop. Wiele z tych rzeczy, które poznaliśmy jest bardzo ważne w pracy jako analitycy danych. Nie możemy się już doczekać kolejnych wyzwań i poszerzania naszej wiedzy.

Spotkania

Spotkania miały formę regularnych warsztatów, podczas których analizowaliśmy oraz wizualizowaliśmy dane. Spotykaliśmy się regularnie co dwa tygodnie, pierwsze spotkania miały początek w listopadzie 2019. Na spotkaniach pracowaliśmy w grupach, zdarzały się również spotkania, gdzie praca była indywidualna. Zawsze nad pomyślnym przebiegiem spotkań czuwali mentorzy.