Politechnika Warszawska

WYDZIAŁ ELEKTRYCZNY

INSTYTUT STEROWANIA I ELEKTRONIKI PRZEMYSŁOWEJ
ZAKŁAD STEROWANIA

Praca dyplomowa inżynierska

na kierunku INFORMATYKA
w specjalności Inżynieria oprogramowania

Predykcja poziomu zapylenia powietrza w Warszawie

Mateusz Gierlach
nr albumu 270746

promotor
dr inż. Grzegorz Sarwas

Warszawa 2018

PREDYKCJA POZIOMU ZAPYLENIA POWIETRZA W WARSZAWIE

Streszczenie

W pracy opisany został problem zapylenia smogowego w obszarach zurbanizowanych oraz został zaproponowany sposób rozwiązania go używając powiązań między zmiennymi pogodowymi i smogowymi. Zrobiony został przegląd literatury dotyczącej dotychczasowych sposobów rozwiązania tego problemu. Zebrane zostały dane pogodowe i ze wskaźników zapylenia. Przeprowadzona została eksploracyjna analiza danych, w celu wydobycia wiedzy na temat powiązań między pogodą, a zapyleniem. Zbudowany został model wyjaśniający te powiązania. Następnie ten model został przetestowany i przedstawione zostały wyniki. W podsumowaniu autor przedstawił możliwe dalsze kierunki rozwoju pracy.

Słowa kluczowe: Eksploracyjna analiza danych, Uczenie maszynowe, Modelowanie statystyczne, Smog, Zapylenie powietrza

PREDICTION OF SMOG AIR POLLUTION IN WARSAW

Abstract

In the thesis, the problem of smog air pollution in urban regions has been described. The author proposed the possible approach to solving it using the relationships between the weather and pollution data. Author researched the solutions that have been published up to this point. Author gathered the data and went through the process of data mining in order to extract the knowledge. The statistical model have been created that explains the relationships in the data. The model has been tested and the effects have been presented. In the summary, the author have outlined the possible next steps in regard to this work.

Keywords: Data mining, Exploratory data analysis, Machine learning, Statistical modelling, Air pollution

Warszawa, 28 lutego 2018

POLITECHNIKA WARSZAWSKA
WYDZIAŁ ELEKTRYCZNY

OŚWIADCZENIE

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa inżynierska pt. Predykcja poziomu zapylenia powietrza w Warszawie:

- została napisana przeze mnie samodzielnie,
- nie narusza niczyich praw autorskich,
- nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam, że przedłożona do obrony praca dyplomowa nie była wcześniej podstawą postępowania związanego z uzyskaniem dyplomu lub tytułu zawodowego w uczelni wyższej. Jestem świadom, że praca zawiera również rezultaty stanowiące własności intelektualne Politechniki Warszawskiej, które nie mogą być udostępniane innym osobom i instytucjom bez zgody Władz Wydziału Elektrycznego.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Mateusz Gierlach.....

Spis treści

Rozdział 1

Wstęp

1.1 Problem smogu

Zanieczyszczenie powietrza jest jednym z największych wyzwań w dzisiejszym świecie. Światowa Organizacja Zdrowia oszacowała [1], że w samym 2012 roku około 7 milionów osób zmarło na skutek chorób powodowanych przez niską jakość powietrza.

Powietrze zanieczyszczane jest zarówno w wyniku naturalnych procesów jak i działalności człowieka. W pracy tej poruszony został temat jakim jest zapylenie smogowe w obszarach zubranizowanych (na przykładzie Warszawy). Jest to jeden z podproblemów związanych z zanieczyszczeniem powietrza, który jest powodowany całkowicie przez działalność człowieka.

Smog to charakterystyczna mgła zawierająca pyły, tlenki azotu, tlenki siarki, pierwiastki metali ciężkich, tlenki węgla, ozon czy węglowodory aromatyczne. Do jego powstania konieczna jest bezwietrzna pogoda.

Wyróżniamy 2 rodzaje smogu:

- smog kwaśny (in. smog londyński, mgła przemysłowa) - występujecy na obszarach miejskich w okresie zimowym. Jego źródłem są przede wszystkim instalacje grzewcze służące do ogrzewania domów, które używają niskiej jakości paliwa.
- smog fotochemiczny (in. smog jasny) - występujący na obszarach miejskich w okresie letnim w dni o dużym nasłonecznieniu. Powstaje na skutek reakcji chemicznej spalin samochodowych i światła słonecznego.

1.2 Smog kwaśny w Polsce

Problem zapylenia smogowego w sezonie zimowym w obszarach miejskich w Polsce jest bardzo poważny - Polska się w czołówce krajów Unii Europejskiej o najbardziej zanieczyszczonym powietrzu. Europejska Agencja Środowiska oszacowała, że w wyniku złej jakości powietrza w Polsce umiera przedwcześnie około 48 tysięcy osób.[2] Oddychanie powietrzem zawierającym smog wywołuje choroby układu oddechowego, w tym także nowotwory płuc, gardła i krtani.

Jedynym typem smogu z jakim mamy do czynienia w Polsce jest smog typu londyńskiego (kwaśny). Najwyższe stężenia pyłów w Polsce odnotowuje się głównie w województwach śląskim (aglomeracja śląska) oraz małopolskim (Kraków). Wysoki poziom smogu jest też obecny w Warszawie, mimo iż to miasto nie leży w dolinie i przepływ wiatru jest w nim większy.

Zdecydowanie najbardziej wpływowym czynnikiem powodującym smog typu kwaśnego w Polsce jest spalanie niskiej jakości paliwa w instalacjach grzewczych w domach jednorodzinnych. W wielu polskich domach wciąż używane są stare instalacje, które spalają dowolny surowiec energetyczny. Bardzo często spalane są w nich węgiel czy odpady, w których zawarte szkodliwe substancje dostają się do atmosfery.

Wpływ na poziom smogu mają również spaliny samochodowe (szacunkowo 15-20%). Dlatego też w centrach miast, gdzie występuje zwiększony ruch uliczny, zapylenie jest zazwyczaj większe.

Wbrew powszechnej opinii o dużym wpływie odpadów przemysłowych na zanieczyszczenie powietrza, jego wpływ w stosunku do powyższych dwóch czynników jest niewielki.

Z racji, iż domy jednorodzinne w klimacie jaki występuje w Polsce są ogrzewane głownie w sezonie zimowym oraz późną jesienią, to właśnie w tym okresie obserwujemy problem zwiększonego smogu.

Poziom zapylenia smogowego może być określany wskaźnikami różnego typu, m.in.: stężenie tlenku siarki SO_2, stężenie tlenku azotu NO_2, stężenie rtęci, zawartość formaldehydu czy stężenie związków ołowiowych.

Jednak wskaźnikami, które w najbardziej ogólny sposób określają poziom zapylenia są:

- PM2,5 - stężenie cząsteczek pyłu zawieszonego o średnicy nie większej niż 2,5 mikrometrów. Są to bardzo małe cząsteczki pyłu wnikające głęboko do płuc i osadzające się na ich pęcherzykach. Określane są jako najbardziej szkodliwe. Są to głównie zanieczyszczenia wtórne z przemian dwutlenku siarki lub azotu.

- PM10 - stężenie cząsteczek pyłu zawieszonego o średnicy nie większej niż 10 mikrometrów. Pyły tego typu osadzają się w drogach oddechowych.Są to głównie cząsteczki metali ciężkich.

1.3 Problem predykcji smogu

Poziom zapylenia smogowego ma dużą zmienność dobową oraz dużą zmienność z dnia na dzień. Poza tym, poziom smogu znacznie różni się między sobą w różnych częściach miasta: w centrum ruch samochodowy jest większy, na dużych osiedlach mieszkalnych wydzielane jest więcej dymu pochodzącego ze spalania. Z powodu tej dużej zmienności w czasie i przestrzeni przewidzenie smogu jest zadaniem niełatwym.

1.4 Możliwe aplikacje praktyczne

Precyzyjnie określona wartość poziomu zapylenia w następnym dniu mogłaby mieć zastosowanie w różnych aplikacjach praktycznych.

Jedną z nich jest system powiadamiania o niebezpieczeństwach. Model użyty w aplikacji, która działa w czasie rzeczywistym pozwalałby na wczesne ostrzeganie o zatruciu powietrza i potencjalną ewakuację ludności z terenów objętych zagrożeniem.

Innym przykładem użycia w praktyce takiego modelu mogłaby być aplikacja, która powiadamia o oczekiwanym zapyleniu powietrza na następny dzień, bazując na prognozie pogody, przez co osoby używające takiej aplikacji miałyby informacje, które pozwoliłyby odpowiedzieć na pytania takie jak: czy wyjście z domu bez maski antysmogowej jest bezpieczne, czy bieganie przy ruchliwej ulicy w następny dzień będzie wiązało się z wdychaniem zanieczyszczonego powietrza itd.

Rozdział 2

Opis pracy inżynierskiej

2.1 Proponowane rozwiązanie problemu predykcji

W tej pracy inżynierskiej, autor postanowił wykonać model predykcyjny poziomu zapylenia smogowego w Warszawie.

Proponowane rozwiązanie oparte jest na następujących założeniach:

- Istnieje zależność między czynnikami pogodowymi oraz poziomem smogu. Stworzony model statystyczny ma opisywać mapowania między przestrzenią danych pogodowych i poziomem zapylenia smogowego w dniu poprzednim, a wektorem opisującym smog w dniu dzisiejszym.
- Typem problemu rozwiązywanym dzięki modelowaniu jest regresja, gdzie zmienną objaśnianą są liczbowe wartości wskaźników PM2,5 i PM10.
- Ogrzewanie mieszkań jest czynnikiem mającym największy wpływ na poziom smogu w miastach i jego analiza jest wystarczająca do określenia poziomu zapylenia. Nie biorę pod uwagi pozostałych czynników takich jak spaliny samochodowe czy odpady przemysłowe.
- Ogrzewanie mieszkań, które jest głównym czynnikiem wywołującym smog, jest zależne od temperatury powietrza - mieszkańcy miast ogrzewają domy tym bardziej im niższa jest temperatura na zewnątrz.
- Prędkość wiatru ma duży wpływ na zapylenie, ponieważ pył zawieszony utrzymuje się przy pogodzie o niskiej wietrzności.
- Znana jest wartość smogu w dniu poprzednim (to już się wydarzyło, mogliśmy zmierzyć) oraz dzisiejsze wskaźniki pogodowe.

- Zapylenie w dniu poprzednim wpływa na dzień dzisiejszy. Mimo dużej zmienności smogu z dnia na dzień, smog nie pojawia się i nie znika w sposób gwałtowny.
- Pogoda w dniu poprzednim wpływa na decyzje mieszkańców miasta o tym czy dalej ogrzewać mieszkanie. Kaloryfery mogą być poodkręcane przez dłuższe okresy czasu i skręcane dopiero, gdy temperatura się znacznie obniży w przeciągu dłuższym niż 1 dzień.
- Predykcja rozumiana jest jako przewidzenie wartości poziomu zapylenia smogowego następnym dniu mając wartość smogu z dnia poprzedniego. Teoretycznie zakładam, że wartości pogodowe dla następnego dnia są znane z krótkoterminowych prognoz pogody, które mają dużą sprawdzalność, a dzisiejszy smog jest już zmierzony. W praktyce przewidzenie smogu nie jest oparte o pobieranie prognoz pogody, a o wywołanie modelu na zbiorze testowym zakładając, że nie znamy wartości smogu w tym dniu.
- Model ma zadanie wyjaśniać jak pogoda wpływa na smog. Głównym celem nie jest uzyskanie jak najlepszego dopasowania modelu do danych, a jego interpretowalność, czyli wyjaśnienie które cechy i w jakim stopniu wpływają na powstawanie smogu.

2.2 Założenia pracy

W pracy autor skupia się na kilku kluczowych podproblemach:

1. Zebranie danych historycznych dotyczących pogody i zapylenia powietrza.
2. Obróbka danych do postaci pozwalającej na analizę.
3. Eksploracyjna analiza danych i znalezienie relacji pomiędzy cechami(zmiennymi).
4. Zbudowanie modelu predykcyjnego z użyciem uczenia maszynowego.
5. Testowanie modelu i interpretacja wyników

Część techniczna została napisana z użyciem języka Python w wersji 3. Zostały użyte następujące biblioteki:

- Pandas - przechowywanie danych
- NumPy - operacje na macierzach
- Matplotlib - wizualizacje danych
- Scikit-learn - algorytmy uczenia maszynowego

Rozdział 3

Analiza istniejących rozwiązań

3.1 Wprowadzenie do przeglądu literatury

Większość komercyjnie stosowanych aplikacji służących do informowania o poziomie zapylenia nie wprowadziło funkcjonalności prognozowania zapylenia na przyszłość. Jest to związane z faktem, że poziom zapylenia jest mocno zmienny nawet na małej przestrzeni geograficznej, a więc do precyzyjnego jego określenia konieczne jest wiele punktów pomiarowych na terenie jednego miasta. Do danych tego typu często nie ma dostępu w czasie rzeczywistym, są udostępniane dopiero po czasie.

W literaturze znane są metody predykcji smogu oparte o znajdowanie zależności między czynnikami pogodowymi, a wskaźnikami zapylenia. Powszechnie używanym podejściem jest modelowanie ekonometryczne dla regresji liniowej i samodzielne dopasowywanie zmiennych do modelu. Coraz częściej spotykanym podejściem w ostatnich czasach jest również użycie algorytmów uczenia maszynowego do automatycznego wyszukiwania powiązań.

W pracy autor wykorzystuje części rozwiązań, które już pojawiły się w innych pracach naukowych. Wprowadzane są też nowe elementy, związane m.in. z inżynierią cech.

W dalszych sekcjach tego rozdziału, autor analizuje dwie prace naukowe, które próbują rozwiązać problem predykcji zanieczyszczenia w niestandardowy sposób oraz porównuje je do zaproponowanego podejścia.

3.2 Air pollution prediction via multi-label classification; G. Corani, M. Scanagatta

W tej pracy[3] rozwiązywanym problemem była klasyfikacja tego czy poziom zanieczyszczenia powietrza na danym obszarze przekroczy pewien ustalony poziom.

Został użyty łączony klasyfikator wielu zmiennych, dzięki któremu możliwe było modelowanie zależności pomiędzy zmiennymi określającymi zanieczyszczenie ozonem, mierzonymi w różnych stacjach. Ta technika pozwoliła uzyskać lepsze wyniki dzięki znalezieniu relacji wielu zmiennych objaśnianych między sobą, które często mają pewną korelację.

W tym celu został napisany klasyfikator wielu zmiennych oparty o sieć Bayesiańską bazującą na uczeniu strukturalnym.

Zmiennymi objaśnianymi były wartości wskaźników ozonu w różnych stacjach. Okazało się, że ów klasyfikator łączony osiągnął lepszą predykcję niż szrego oddzielnych modeli dla każdej stacji pomiarowej. Zostały również poznane zależności między pomiarami na stacjach.

Podejście w proponowanym rozwiązaniu w stosunku do zaprezentowanego tutaj:

- Prognozowana jest konkretna wartość poziomu smogu, a nie to czy zostanie przekroczona pewna granica
- Założone jest, że wartości smogu w punktach pomiarowych są od siebie niezależne, korelacja między nimi nie jest poszukiwana
- Model oparty jest o drzewa losowe, nie o sieć Bayesiańską

3.3 Data mining methods for prediction of air pollution; K. Siwek, S. Osowski

W tej pracy[4] zostały przedstawione techniki eksploracji danych służące predykcji zanieczyszczenia powietrza. W pierwszej części został opisany problem generacji i wyboru zmiennych objaśniających, w drugiej części system prognozujący zanieczyszczenie na następny dzień.

Jako zmienne objaśniające zostały wybrane parametry atmosferyczne. Zmiennymi objaśnianymi są wartości PM10, SO2 (tlenek siarki), NO2 (tlenki azotu) i O3 (ozon).

Użyte są 2 techniki służące do wyboru optymalnych zmiennych do modelu: podejście globalne oparte o analizę całego zbioru danych oraz liniowa metoda stopniowego dopasowania do danych.

Przedstawione są również 2 podejścia do modelowania zależności między zmiennymi meteorologicznymi, a smogowymi. Pierwszą z nich są lasy losowe jako ensembling drzew decyzyjnych, drugą są predyktory oparte o sieć neuronową. Ostatecznie łączone są oba modele.

Praca wykazuje, że dzięki użyciu algorytmów optymalizujących do wyboru cech modelu oraz łączenia podejść tradycyjnego (RandomForest) i sieci neuronowych, możemy uzyskać znacznie wydajności predykcji zanieczyszczenia powietrza.

Podejście w proponowanym rozwiązaniu w stosunku do zaprezentowanego tutaj:

- Algorytm wybierający cechy nie jest używany, a zamiast tego przeprowadzana jest eksploracyjna analizę zależności w danych, aby samodzielnie wybrać je do modelu
- Nie występuje łączenie modeli lasów losowych z siecią neuronową, ponieważ bardziej interesująca jest interpretowalność wyników i wyciągnięcie logicznych wniosków, a nie jak najwyższa predykcja

Rozdział 4

Zebranie i przygotowanie danych

4.1 Zebranie danych pogodowych

Dokładne dane pogodowe ze stacji mierniczych są udostępniane przez Instytut Meteorologii i Gospodarki Wodnej na ich stronie: dane.imgw.pl.

Pobrane zostały 2 paczki synoptycznych danych pogodowych dobowych dla zakresu czasowego od 1.01.2011 do 31.12.2017 zmierzone na stacji badawczej nr 352200375 w Warszawie.

Lokalizacja stacji badawczej:

Rysunek 4.1: Położenie stacji badawczej dla danych pogodowych

1 paczka zawierała następujące zmienne:

- Średnie dobowe zachmurzenie ogólne
- Status pomiaru NOS
- Średnia dobowa prędkość wiatru
- Status pomiaru FWS
- Średnia dobowa temperatura
- Status pomiaru TEMP
- Średnia dobowe ciśnienie pary wodnej
- Status pomiaru CPW
- Średnia dobowa wilgotność względna
- Status pomiaru WLGS
- Średnia dobowe ciśnienie na poziomie stacji
- Status pomiaru PPPS
- Średnie dobowe ciśnienie na pozimie morza
- Status pomiaru PPPM
- Suma opadu dzień
- Status pomiaru WODZ
- Suma opadu noc
- Status pomiaru WONO

2 paczka zawierała następujące zmienne:

- Maksymalna temperatura dobowa
- Status pomiaru TMAX
- Minimalna temperatura dobowa
- Status pomiaru TMIN
- Średnia temperatura dobowa
- Status pomiaru STD
- Temperatura minimalna przy gruncie
- Status pomiaru TMNG
- Suma dobowa opadu

- Status pomiaru SMDB
- Rodzaj opadu
- Wysokość pokrywy śnieżnej
- Status pomiaru PKSN
- Równoważnik wodny śniegu
- Status pomiaru RWSN
- Usłonecznienie
- Status pomiaru USL
- Czas trwania opadu deszczu
- Status pomiaru DESZ
- Czas trwania opadu śniegu
- Status pomiaru SNEG
- Czas trwania opadu deszczu ze śniegiem
- Status pomiaru DISN
- Czas trwania gradu
- Status pomiaru GRAD
- Czas trwania mgły
- Status pomiaru MGLA
- Czas trwania zamglenia
- Status pomiaru ZMGL
- Czas trwania sadzi
- Status pomiaru SADZ
- Czas trwania gołoledzi
- Status pomiaru GOLO
- Czas trwania zamieci śnieżnej niskiej
- Status pomiaru ZMNI
- Czas trwania zamieci śnieżnej wysokiej
- Status pomiaru ZMWS
- Czas trwania zmętnienia
- Status pomiaru ZMET
- Czas trwania wiatru >=10m/s

- Status pomiaru FF10
- Czas trwania wiatru >15m/s
- Status pomiaru FF15
- Czas trwania burzy
- Status pomiaru BRZA
- Czas trwania rosy
- Status pomiaru ROSA
- Czas trwania szronu
- Status pomiaru SZRO
- Wystąpienie pokrywy śnieżnej
- Status pomiaru DZPS
- Wystąpienie błyskawicy
- Status pomiaru DZBL
- Stan gruntu Z/R
- Izoterma dolna
- Status pomiaru IZD
- Izoterma górna
- Status pomiaru IZG
- Aktynometria
- Status pomiaru AKTN

4.2 Zebranie danych smogowych

Dane ze wskaźników zapylenia smogowego są udostępniane przez Główny Inspektorat Ochrony Środowiska na stronie: http://powietrze.gios.gov.pl.

Dla zakresu czasowego takiego jak dla danych pogodowych, odczyty były dostępne dla 2 punktów pomiarowych na terenie Warszawy:

- Wokalna na Ursynowie
- Róg Aleji Niepodległości i Nowowiejskiej na Śródmieściu

Pobrałem dane smogowe (PM2,5, PM10) godzinne dla zakresu czasowego od 1.01.2011 do 31.12.2016. Dane dla 2017 roku nie zostały jeszcze udostępnione.

Zmienne smogowe:

- PM2,5 dla Wokalnej
- PM10 dla Wokalnej
- PM2,5 dla Niepodległości
- PM10 dla Niepodległości

Lokalizacje punktów pomiarowych:

Rysunek 4.2: Położenie punktów pomiarowych dla danych zapylenia smogowego

4.3 Obróbka wstępna danych pogodowych

Z wielu zmiennych, które są w paczkach wybrane zostały tylko te, które mają szanse wyjaśniać smog w jakikolwiek sposób. Kolejnym krokiem jest odrzucenie z 2 paczki zmiennych powtarzających się z 1. Poza tym usunięte

zostały zmienne statusowe. Daty rozbite zostały na osobne kolumny: rok, miesiąc, dzień.

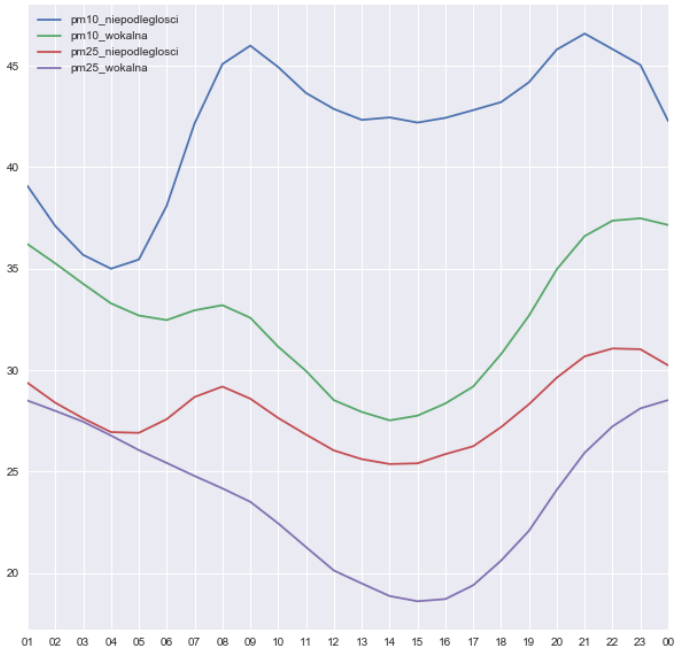
Następnie łączone są obie zmienne i usuwane rekordy dla roku 2017, ponieważ nie ma danych dla tego zakresu w przypadku zapylenia PM2,5 i PM10.

Po wstępnym przygotowaniu w ramce danych pogodowych zostają następujące zmienne:

- rok
- miesiac
- dzien
- max_temp - maksymalna temperatura zarejestrowana danego dnia
- min_temp - minimalna temperatura zarejestrowana danego dnia
- sr_temp - średnia temperatura w danym dniu
- min_temp_grunt - minimalna temperatura przy gruncie zarejestrowana danego dnia
- suma_opad - suma opadów w danym dniu
- rodz_opad - rodzaj opadów; W - deszcz, S - śnieg
- sr_zachm - średnie zachmurzenie w danym dniu
- sr_predk_wiatr - średnia prędkość wiatru w danym dniu
- sr_cisn - średnie dobowe ciśnienie pary wodnej
- sr_wilg - średnia wilgotność w danym dniu
- sr_cisn_stacja - średnie ciśnienie atmosferyczne na poziomie stacji pomiarowej
- sr_cisn_morze - średnie ciśnienie atmosferyczne na poziomie stacji morza
- opad_dzien - suma opadów za dnia (nie podano jak określany dzień/noc)
- opad_noc - suma opadów w nocy

4.4 Obróbka wstępna danych smogowych

Dane smogowe są godzinne - po 24 rekordy dla każdego dnia. Sprawdzane jest jak wyglądają średnie wartości odczytów smogowych dla obu stacji w zależności od godziny:



Rysunek 4.3: Średnie wartości smogu / godzina dnia

Zauważmy, że istnieje lokalne maksimum dla godziny 8 rano. Jest to specyficzne miejsce, ponieważ wtedy znaczna część osób mieszkańców Warszawy jedzie samochodami do pracy. Precyzyjne wyjaśnienie wartości w tym miejscu może być ciężkie właśnie ze względu na dodatkowy aspekt zwiększonego ruchu ulicznego w tym czasie, jednak zbudowanie modelu również dla precyzyjnego określenia smogu właśnie dla godziny 8 jest istotne, ponieważ z punktu widzenia praktycznego jest to godzina, dla której chcemy znać dokładnie poziom zapylenia, bo właśnie wtedy wiele osób przemieszcza się po mieście i są szczególnie narażeni na negatywne działanie smogu.

A więc agregując dane godzinne na dni tworzone są zmienne określające wartości PM2,5 i PM10 dla obu punktów pomiarowych (Niepodległości, Wokalna) dla 2 przypadków: wartości średnie z całej doby oraz wartości konkretne dla godziny 8 rano.

Po wstępnym przygotowaniu w ramce danych smogowych zostają następujące zmienne:

- rok
- miesiac
- dzien
- pm10_sr_niep - średnia dobowa wartość PM10 na Niepodległości
- pm10_sr_wok - średnia dobowa wartość PM10 na Wokalnej
- pm25_sr_niep - średnia dobowa wartość PM2,5 na Niepodległości
- pm25_sr_wok - średnia dobowa wartość PM2,5 na Wokalnej
- pm10_8_niep - wartość PM10 na Niepodległości o 8 rano
- pm10_8_wok - wartość PM10 na Wokalnej o 8 rano
- pm25_8_niep - wartość PM2,5 na Niepodległości o 8 rano
- pm25_8_wok - wartość PM2,5 na Wokalnej o 8 rano

Rozdział 5

Eksploracyjna analiza danych

5.1 Połączenie danych, brakujące wartości

Kolejnym krokiem jest sprowadzenie danych do jednej ramki. Łączone są obie ramki po zmiennych czasowych (rok, miesiac, dzien).

Wejściowa ramka danych, na której przeprowadzana jest eksploracja ma 2192 rzędy reprezentujące kolejne dni od 1.01.2011 do 31.12.2016 oraz 25 kolumn reprezentujące zmienne kalendarzowe, pogodowe i smogowe.

Typy i liczności zmiennych:
rok 2192 non-null int64
miesiac 2192 non-null int64
dzien 2192 non-null int64
max_temp 2192 non-null float64
min_temp 2192 non-null float64
sr_temp 2192 non-null float64
min_temp_grunt 2192 non-null float64
suma_opad 2192 non-null float64
rodz_opad 1274 non-null object
sr_zachm 2192 non-null float64
sr_predk_wiatr 2192 non-null float64
sr_cisn 2192 non-null float64
sr_wilg 2192 non-null float64
sr_cisn_stacja 2192 non-null float64
sr_cisn_morze 2192 non-null float64
opad_dzien 2192 non-null float64
opad_noc 2192 non-null float64
pm10_sr_niep 2078 non-null float64
pm10_sr_wok 2175 non-null float64

pm25_sr_niep 2083 non-null float64
pm25_sr_wok 2176 non-null float64
pm10_8_niep 2049 non-null float64
pm10_8_wok 2144 non-null float64
pm25_8_niep 2053 non-null float64
pm25_8_wok 2148 non-null float64
Zauważmy:

- rok, miesiac, dzien - typ integer, ale określają klasy - zamieniam na typ kategoryczny
- brakujące wartości tylko dla zmiennych smogowych oraz rodz_opad
- rodz_opad ma aż 918 brakujących wartości - usuwam kolumnę

Następnie analizowane są miejsca, w których występują brakujące wartości dla zmiennych smogowych. Okazuje się, że są to w większości te same rekordy dla różnych zmiennych smogowych. Ponadto zdecydowana większość z nich to brakujące wartości przez wiele rekordów z rzędu, a więc nie ma jak ich zastąpić np. średnią z okolicznych dni. Rekordy z brakującymi wartościami zmiennych smogowych to około 9,5 % wszystkich rekordów. Zostają usunięte, ponieważ nie spowoduje to drastycznego obniżenia ilości danych.

5.2 Korelacje zmiennych

Niektóre zmienne opisują podobne czynniki. Złą praktyką jest wkładanie do modelu wielu czynników reprezentujących te same informacje pogodowe. Do sprawdzania podobieństw między zmiennymi używana jest korelacji Pearsona. Sprawdzane są korelacje między sobą podobnych zmiennych.

Poniżej przedstawione jest jak prezentują się korelacje między zmiennymi w kilku grupach tematycznych: temperatury, opady, ciśnienia oraz powiązanie opadów z wilgotnością.

	max_temp	min_temp	sr_temp	min_temp_grunt
max_temp	1.000000	0.918868	0.985219	0.871570
min_temp	0.918868	1.000000	0.962363	0.986598
sr_temp	0.985219	0.962363	1.000000	0.925086
min_temp_grunt	0.871570	0.986598	0.925086	1.000000

Rysunek 5.1: Korelacje między temperaturami

	suma_opad	opad_dzien	opad_noc
suma_opad	1.000000	0.800668	0.729888
opad_dzien	0.800668	1.000000	0.174867
opad_noc	0.729888	0.174867	1.000000

Rysunek 5.2: Korelacje między opadami

	sr_cisn	sr_cisn_stacja	sr_cisn_morze
sr_cisn	1.000000	-0.197265	-0.238848
sr_cisn_stacja	-0.197265	1.000000	0.998831
sr_cisn_morze	-0.238848	0.998831	1.000000

Rysunek 5.3: Korelacje między ciśnieniami

	suma_opad	sr_wilg
suma_opad	1.000000	0.212625
sr_wilg	0.212625	1.000000

Rysunek 5.4: Korelacja między opadami, a wilgotnością

Wyciągane są następujące wnioski:

- Temperatury są mocno skorelowane ze sobą - zostawiane są tylko temperaturę średnią, która najlepiej reprezentuje ten czynnik, pozostałe usuwam
- Opady w dzień i w nocy nie są skorelowane, ale suma opadów dobrze reprezentuje całość zależności opadów - zostawiane są tylko sumę opadów, pozostałe usuwam
- Ciśnienia na poziomie stacji i morza są mocno skorelowane. Zostawiane są ciśnienie na poziomie morza, ponieważ taka informacja jest łatwiejsza do wyciągnięcia np. z prognozy pogody, standardowo ciśnienie podaje się w ten sposób, więc jest to lepiej interpretowalne. Dodatkowo

usuwane są średnie ciśnienie pary wodnej, ponieważ w żaden sposób nie powinno wpływać to na smog.

- Opady i wilgotność powietrza nie są skorelowane. Zostawiane są obie.

Następnie sprawdzana jest korelacja Pearsona między zmiennymi objaśniającymi (pogodą), a objaśnianymi (smogiem). Nie ma tutaj żadnych skorelowanych zmiennych.

Kolejnym krokiem jest sprawdzenie korelacji między zmiennymi objaśnianymi, aby sprawdzić czy smog można reprezentować za pomocą mniejszej liczby zmiennych.

	pm10_sr_niep	pm10_sr_wok	pm25_sr_niep	pm25_sr_wok	pm10_8_niep	pm10_8_wok	pm25_8_niep	pm25_8_wok
pm10_sr_niep	1.000000	0.830792	0.869079	0.777108	0.835002	0.749995	0.822356	0.740254
pm10_sr_wok	0.830792	1.000000	0.901357	0.938286	0.657073	0.880609	0.830645	0.884862
pm25_sr_niep	0.869079	0.901357	1.000000	0.926419	0.653508	0.770606	0.895331	0.836828
pm25_sr_wok	0.777108	0.938286	0.926419	1.000000	0.563977	0.803185	0.826347	0.912383
pm10_8_niep	0.835002	0.657073	0.653508	0.563977	1.000000	0.741019	0.803112	0.655416
pm10_8_wok	0.749995	0.880609	0.770606	0.803185	0.741019	1.000000	0.845231	0.908337
pm25_8_niep	0.822356	0.830645	0.895331	0.826347	0.803112	0.845231	1.000000	0.862118
pm25_8_wok	0.740254	0.884862	0.836828	0.912383	0.655416	0.908337	0.862118	1.000000

Rysunek 5.5: Korelacja między zmiennymi objaśnianymi - smogowymi

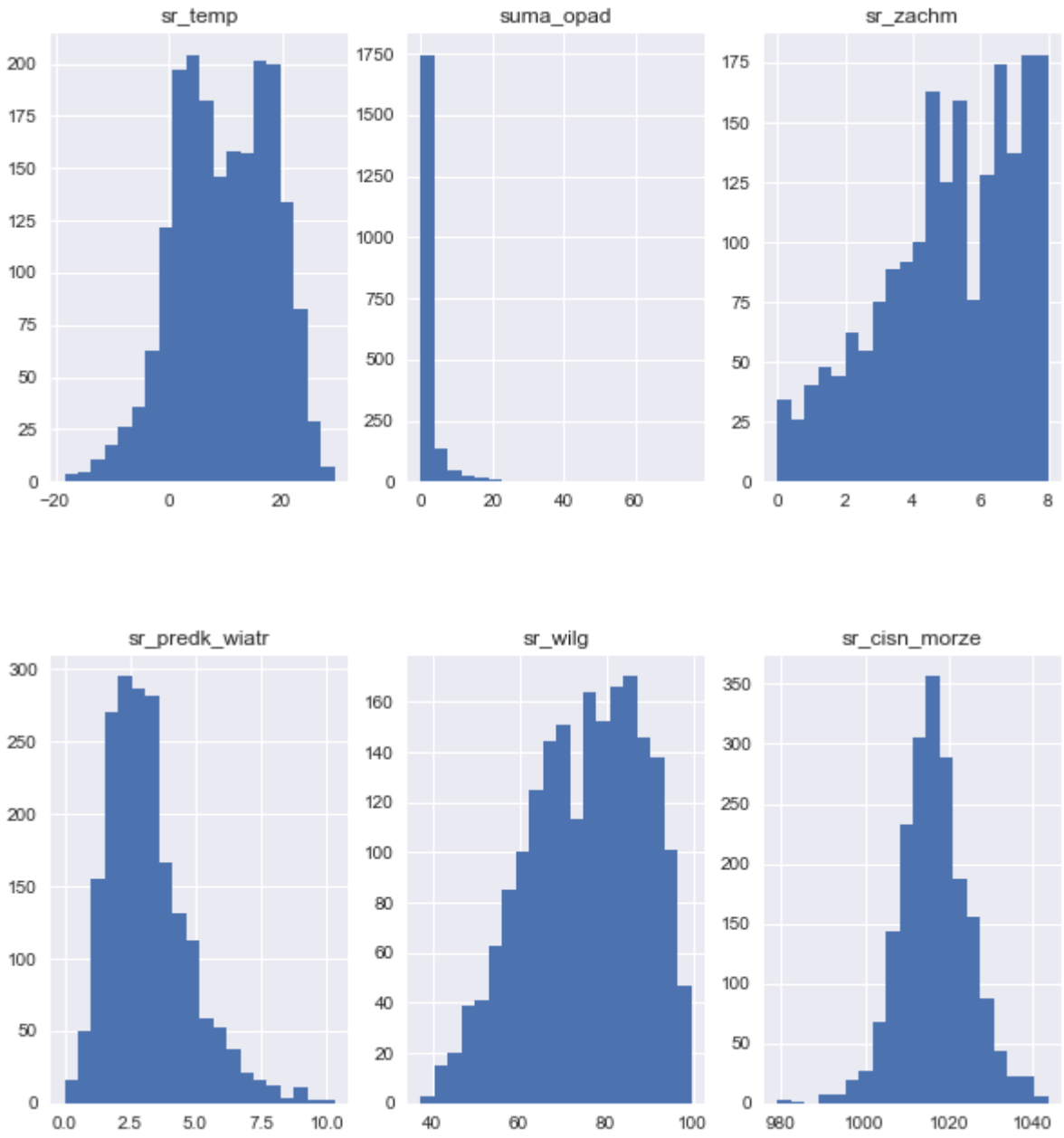
Zauważmy:

- Zmienna pm25_8_niep najlepiej reprezentuje dane - jej korelacja z każdą z innych zmiennych nigdy nie jest mniejsza niż 80
- pm10_8_niep i pm25_sr_wok są bardzo słabo skorelowane. Istnieje obawa, że pm25_8_niep niewystarczająco będzie rperezentować różnice w zależnościach między nimi
- Ostatecznie: zostawiam 3 zmienne smogowe: pm25_8_niep, pm10_8_niep, pm25_sr_wok - będę robił 3 osobne modele dla tych 3 wskaźników

5.3 Wartości odstające

Autor zaczyna od sprawdzenia wartości odstających dla zmiennych objaśniających. Używa w tym celu histogramów, posiłkując się dodatkowo sprawdzaniem ilości wierszy powyżej jakiejś wartości dla konkretnej analizowanej zmiennej.

Histogramy dla zmiennych pogodowych:

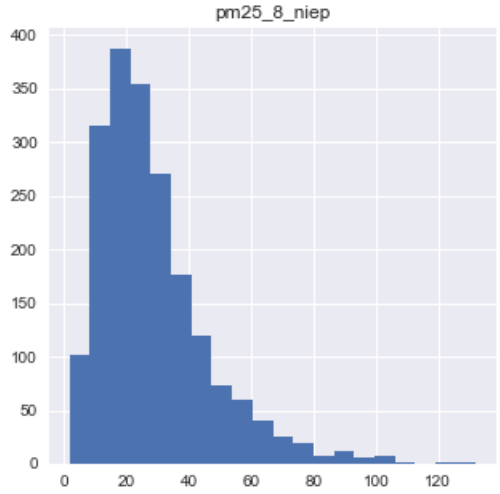


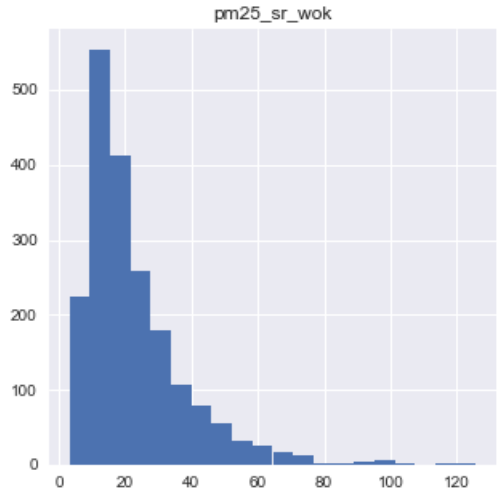
Rysunek 5.6: Histogramy dla zmiennych objaśniających

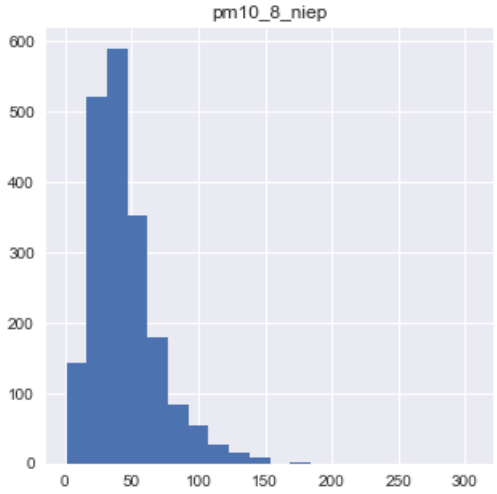
Zauważmy, że wartości wyraźnie odstające od trendu są tylko w sumie opadów. Usuwane są rekordy, gdzie wartość tej kolumny jest większa niż 36.

Kolejnym krokiem jest sprawdzenie czy nie występują wartości odstające w zmiennych objaśnianych (zmienne smogowe).

Histogramy dla zmiennych smogowych:







Zauważmy:

- pm25_8_niep ma kilka wartości odstających. Usuwam rekordy dla których zmienna większa od 150.
- pm25_sr_wok nie ma wartości odstających.
- pm10_8_niep ma wartości odstające, Usuwam rekordy dla których zmienna jest większa pd 180

5.4 Inżynieria cech

Oprócz wartości liczbowych zmiennych, ważne jest również by przedstawić w danych pewne dodatkowe aspekty, które mogą polepszyć wyjaśnialność modelu. Przeprowadzany zostaje proces inżynierii cech, czyli tworzenia nowych cech bazując na tych już istniejących.

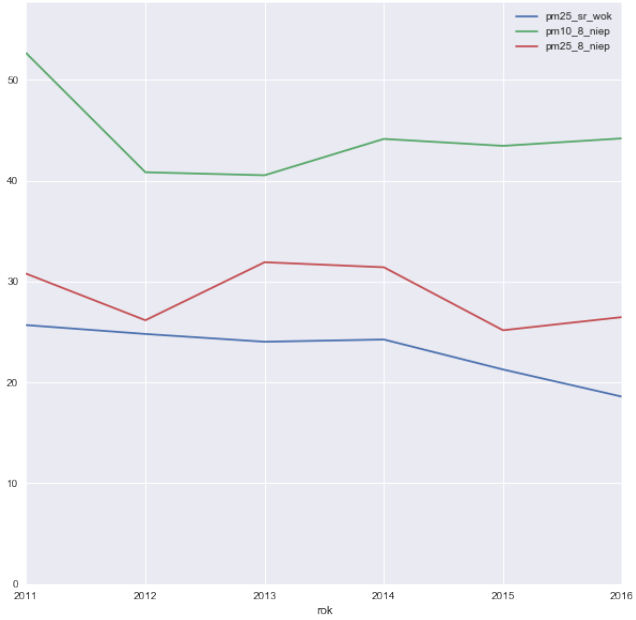
Nowe zmienne:

- bez_wiatru to zmienna boolowska przyjmująca wartość True, jeśli w danym dniu predkosc wiatru była równa 0 i False jeśli różna od 0
- poprz_sr_temp to zmienna przechowująca średnią temperaturę z dnia poprzedniego, ponieważ temperatura z dnia wcześniejszego wpływa na to, czy ludzie dalej grzeją w kaloryferach/piecach

- Ważne są również zmiany czynników pogodowych z dnia na dzień. To jak zmieniły się czynniki pogodowe z wczoraj na dzisiaj ma wpływ na zachowanie ludzi, jeśli chodzi o ogrzewanie domu. Tak więc tworzony jest szereg nowych zmiennych które są obliczane jako różnica między wartością czynnika z dzisiaj, a wartością z wczoraj. Tworzone są one dla tych czynników: suma opadów, zachmurzenie, prędkość wiatru, wilgotność, ciśnienie atmosferyczne.

5.5 Smog na przestrzeni lat i miesięcy

Ważne, by wiedzieć jaki jest trend wartości zapylenia smogowego w zależności od kolejnych lat. Należy sprawdzić czy mam jakiś trend długoterminowy, którego nie dałoby się opisać za pomocą samej pogody i smogu z poprzedniego dnia. Sprawdzany jest on za pomocą wyliczenia średnich wartości zapylenia smogowego dla 3 zmiennych objaśnianych na przestrzeni lat.

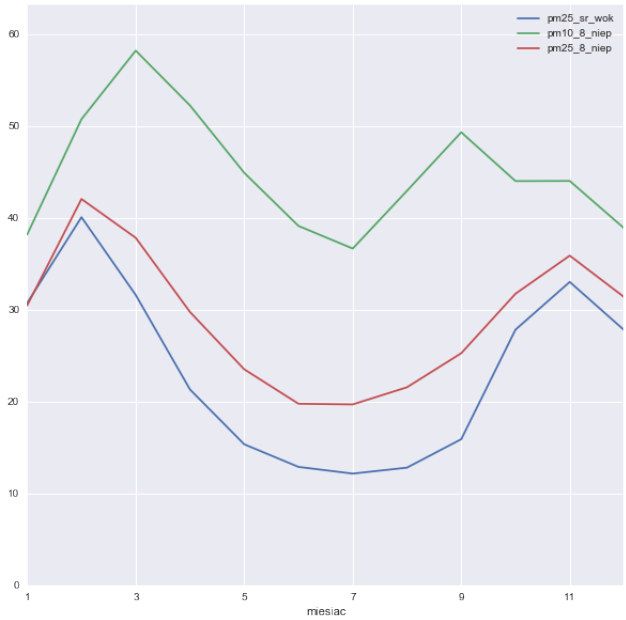


Rysunek 5.7: Średnie zapylenie na przestrzeni lat

Zauważmy, że nie ma widocznych trendów zmiennych smogowych na przestrzeni lat.

Zobaczmy jak wyglądają uśrednione wartości smogowe na przestrzeni miesięcy. Oczywiste jest, że smog jest duży w sezonie zimowym, ale warto

przeanalizować jak to wygląda na wykresie.

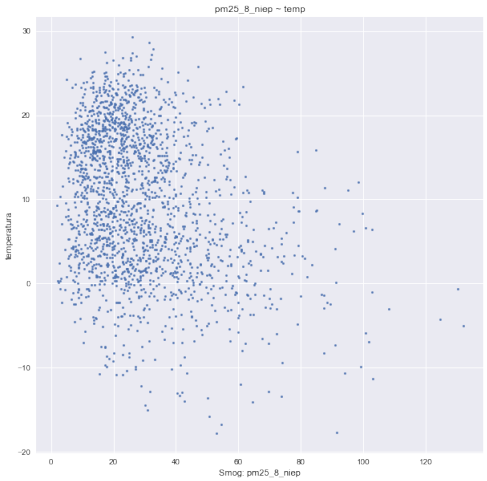


Rysunek 5.8: Średnie zapylenie na przestrzeni miesięcy

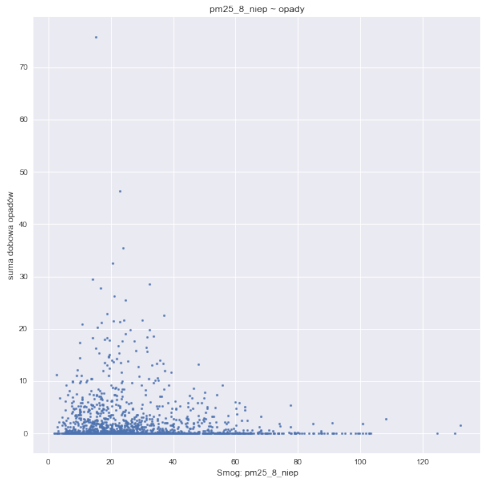
Zauważmy, że duże zapylenie jest od października do marca. Zmienna kategoryczna 'miesiac' może w pewien sposób pomagać w predykcji. Autor zostawia ją w ramce i będzie brał ją do modelu.

5.6 Zależności między zmiennymi objaśniającymi, a objaśnianymi

Spodziewamy się, że zmienne objaśniające będą układały się w pewne zależności ze zmiennymi objaśnianymi, np. spodziewane jest, że wraz ze spadkiem temperatury, poziom smogu będzie się zwiększał. Przeanalizujmy wykresy punktowe i sprawdźmy czy rzeczywiście występują jakieś konkretne zależności. Jako zmienną reprezentującą smog wybieram PM2,5 o 8 na Niepodległości, ponieważ jest ona wysoko skorelowana ze wszystkimi innymi wskaźnikami zapylenia.

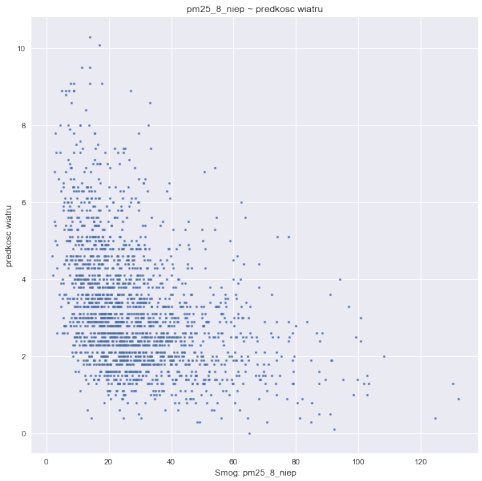


Rysunek 5.9: Smog - średnia temperatura

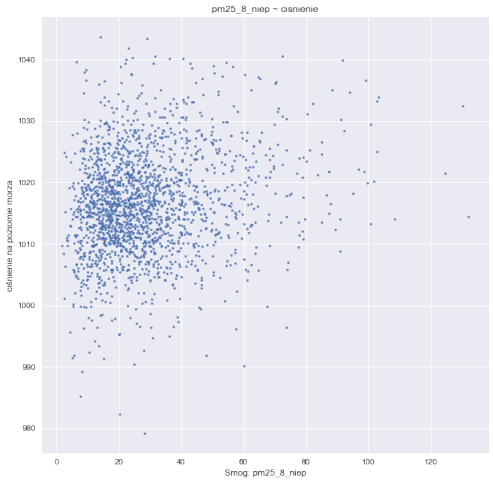


Rysunek 5.10: Smog - dobowa suma opadów

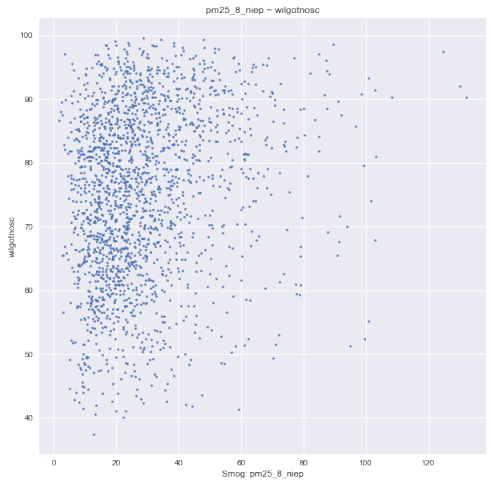
Rysunek 5.11: Smog - średnie zachmurzenie



Rysunek 5.12: Smog - średnia prędkość wiatru



Rysunek 5.13: Smog - średnie ciśnienie na poz. morza



Rysunek 5.14: Smog - średnia wilgotność

Spostrzeżenia:

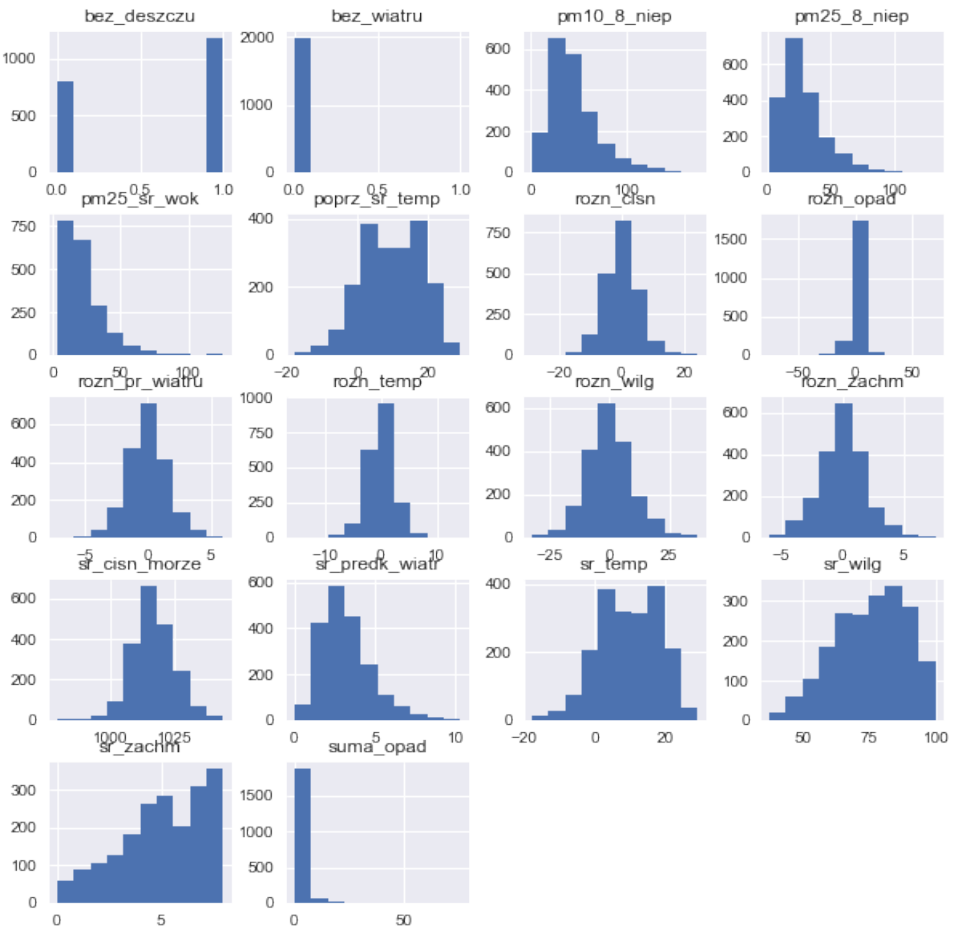
- Spadek temperatury związany ze wzrostem zapylenia - niewyraźny trend
- Brak widocznej zależności między opadami, a smogiem
- Brak widocznej zależności między zachmurzeniem, a smogiem
- Wraz ze wzrostem prędkości wiatru, smog jest coraz mniejszy - wyraźny trend

- Wraz ze wzrostem ciśnienia atmosferycznego, smog również się zwiększa - średni trend
- Przy większej wilgotności, smog również zwiększa się w niewielkim stopniu - bardzo słaby trend

5.7 Skośność rozkładu zmiennych

Mimo iż algorytm lasów losowych, który jest używany do modelowania dobrze radzi sobie z nierównościami w rozkładzie danych (m.in. dlatego nie przeprowadzam regularyzacji czy standaryzacji zmiennych), to warto "wyrównaćżozkłady niektórych zmiennych, aby zmniejszyć ważność małych wartości i wystarczająca reprezentować też duże wartości zmiennej.

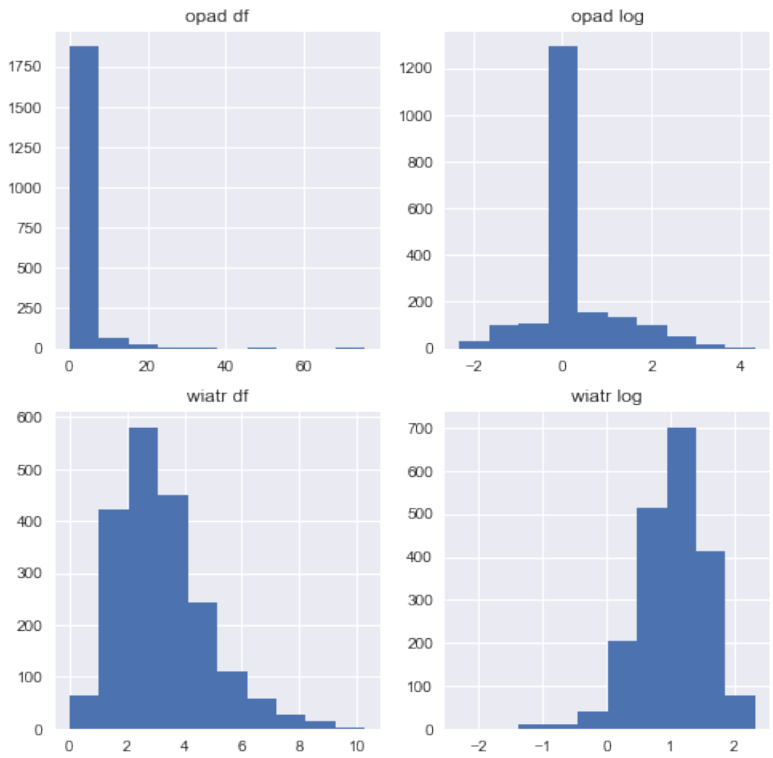
Skośności możemy w łatwy sposób zobaczyć rysując histogramy. Zobaczmy jak wyglądają histogramy wszystkich zmiennych:



Rysunek 5.15: Histogramy zmiennych

Zmienne objaśniane o rozkładzie prawoskośnym to suma_opad oraz sr_predk_5wiatr. Logarytmuję je, aby wyrównać rozkłady. Jednak te 2 zmienne zlogarytmowane zostaną podmienione w nowej ramce. Będzie testowane modelowanie na 2 ramkach: ze zmiennymi objaśnianymi zlogarytmowanymi i bez wykonywania tej operacji.

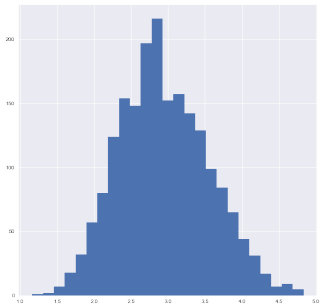
Różnice między rozkładem dla opadów i prędkości wiatru przed i po zlogarytmowaniu:



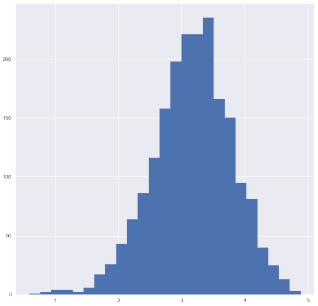
Rysunek 5.16: Opady i prędkość wiatru przed i po zlogarytmowaniu

Następnie autor przechodzi do zmiennych smogowych. Tutaj wszystkie 3 zmienne są wyraźnie prawoskośne. Tutaj również stosowane jest podejście związane z logarytmowaniem i podziale na 2 części (zlogarytmowane, niezlogarytmowane), które będa testowane osobno w modelowaniu.

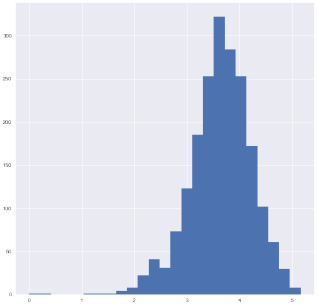
Tak wyglądają zmienne smogowe po zlogarytmowaniu:



Rysunek 5.17: PM2,5 średnie Wokalna - po zlogarytmowaniu



Rysunek 5.18: PM2,5 o 8 Niepodległości - po zlogarytmowaniu



Rysunek 5.19: PM10 o 8 Niepodległości - po zlogarytmowaniu

Możemy zauważyć wartości odstające w przebiegach. Dopasowywane są do najmniejszej wartości w histogramie, aby na tym etapie nie usuwać zmiennych, bo uniemożliwiło by to porównanie modeli zbudowanych na 2 ramkach.

Tak obrobione dane są gotowe do modelowania.

Rozdział 6

Modelowanie

Modelowanie jest ostatnim krokiem na drodze do znalezienia powiązań między czynnikami pogodowymi i smogiem. Właściwie zbudowany model jest narzędziem służącym do wyjaśnienia tych powiązań. Sukces etapu modelowania zależy od wcześniejszego poprawnego obrobienia danych, wyboru optymalnego algorytmu do rozwiązywanego problemu oraz umiejętności wyciągnięcia poprawnych wniosków. W tym rozdziale, autor opisuje kolejne etapy składające w procesie modelowania.

6.1 Algorytm lasów losowych w problemie regresji

Do tworzenia modelu wybrany został algorytm lasów losowych dla regresji. Radzi sobie on dobrze zarówno ze zmiennymi ciągłymi jak i kategorycznymi. Nie wymaga skalowania zmiennych w żaden sposób (regularyzacja, standaryzacja itd.) i dzięki wyliczaniu ważności zmiennych daje narzędzie do interpretowanie modelu - ważności zmiennych mówią o tym jak bardzo dana zmienna wpływa na określenie predykcji modelu i pokazuje, które zmienne są najistotniejsze. Lasy losowe są przykładem algorytmu ensemblingowego, który składa się z wielu drzew decyzyjnych. Ilość drzew używanych w modelu jest jedynym modyfikowalnym parametrem.

Poza tym jego dokładność jest bardzo dobra w stosunku do innych tradycyjnych algorytmów uczenia maszynowego. Oczywiście jest mniej dokładny niż np. sieci neuronowe, ale w przypadku tego modelu ważne jest zarówno stosunkowo niezłe dopasowanie jak i wyjaśnialność, czego sieć neuronowa nie zapewnia.

6.2 Wstęp do procesu modelowania

Zanim będą wyciągane wnioski dotyczące całego problemu przewidzenia smogu, autor chce najpierw dzięki procesowi modelowania znaleźć odpowiedzi na następujące pytania:

- Czy lepsze jest uczenie na zbiorze ze zmiennymi objaśniającymi zlogarytmowanymi czy na standardowych?
- Czy lepiej, aby zmienna objaśniana była zlogarytmowana czy nie? Czy same czynniki pogodowe są wystarczające, aby dobrze określić predykcję, czy objaśnianie za pomocą smogu w dniu poprzednim jest absolutnie konieczne?

Ostatecznie dzięki modelowi autor chce znaleźć odpowiedzi na pytania dotyczące całego procesu predykcji zapylenia smogowego:

- Które czynniki pogodowe są najważniejsze?
- Czy smog w różnych punktów Warszawy możemy przewidzieć z podobną precyzją?
- Czy smog o 8 rano dobrze możemy przewidzieć z podobną predykcją jak uśredniony smog dla całego dnia?

Za każdym razem, gdy tworzone są modele, dokonywany jest podział danych na zbiór treningowy i testowy w stosunku 80/20. Modele między sobą są porównywane za pomocą 3 statystyk opisujących zależności między wartościami rzeczywistymi z mierników smogu, które nie zostały użyte do tworzenia modelu (20%), a predykcją modelu na danych testowych:

- Współczynnik determinacji r-kwadrat (R2) opisujący miarę dopasowania modelu (jest to najważniejsza miara)
- Korelacja Spearmana
- Korelacja Pearsona

6.3 Model oparty na samej pogodzie - kwestia logarytmowania

W pierwszej kolejności tworzony jest model, który opisuje zależność między samymi czynnikami pogodowymi (nie biorąc pod uwagę smogu w dniu poprzednim).

Jest on tworzony po to, aby zobaczyć jak zmienia się predykcja, gdy 2 z czynników pogodowych są logarytmowane. Wykonywane są modele tylko dla 1 zmiennej objaśnianej: pm25_8_niep. Las losowy tworzony jest w oparciu o 500 drzew decyzyjnych.

Zmienne pogodowe bez logarytmowania:

R2 na zb. testowym: 0.485 Kor. Spearmana na zb. testowym: 0.704 Kor. Pearsona na zb. testowym: 0.701

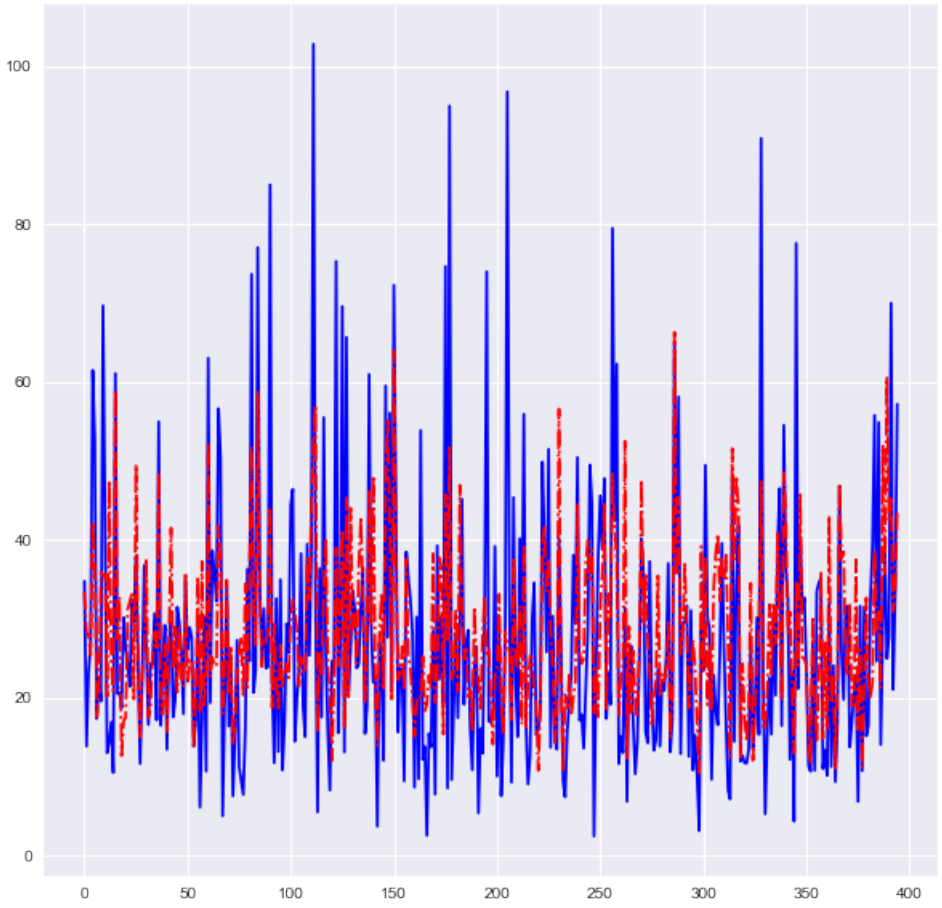
Zmienne pogodowe z logarytmowaniem:

R2 na zb. testowym: 0.488 Kor. Spearmana na zb. testowym: 0.673 Kor. Pearsona na zb. testowym: 0.705

Jak widzimy, logarytmowanie zmiennych objaśniających w prawie żaden sposób nie wpływa na polepszenie modelu. W dalszej części modelowania, używane są tylko ramki danych oparte o standardowe wartości zmiennych pogodowych.

Zauważmy, że poziom predykcji, gdy smog próbujemy objaśniać samą pogodą jest słaby. Pokrycie wartości testowych na poziomie ok. 49% jest niewystarczające do precyzyjnego przewidzenia wartości smogu.

Na wykresie dokładnie widać jak słabe jest pokrycie (niebieski - wartości prawdziwe, czerwona przerywana - predykcja modelu):



Rysunek 6.1: Pokrycie modelu bazującego na samych zmiennych pogodowych

6.4 Model oparty na pogodzie i smogu z dnia poprzedniego - kwestia logarytmowania

Predykcja smogu oparta na czynnikach pogodowych oraz smogu z dnia poprzednie ma duży sens, ponieważ zakładamy, że smog historyczny jest znany.

Bardzo istotne jest to, że smog z dnia poprzedniego nie jest czynnikiem wyjaśniającym to jaki poziom smogu jest dzisiaj, wyjaśniamy to tylko za pomocą zmiennych pogodowych. Ale wiedząc, że smog nie zmienia się w sposób bardzo gwałtowny, nasza predykcja powinna być dokładniejsza, gdy znamy bazę z dnia poprzedniego, a pogodą staramy się wyjaśnić zmianę zapylenia.

W tym kroku autor chce sprawdzić czy logarytmowanie zmiennych smogowych (zarówno zmienna objaśniana jak i objaśniająca smogowa dla dnia poprzedniego) daje polepszenie predykcji. Znów wykonywane modele tylko dla 1 zmiennej objaśnianej: pm25_8_niep.

Zmienne pogodowe bez logarytmowania:

R2 na zb. testowym: 0.54 Kor. Spearmana na zb. testowym: 0.71 Kor. Pearsona na zb. testowym: 0.737

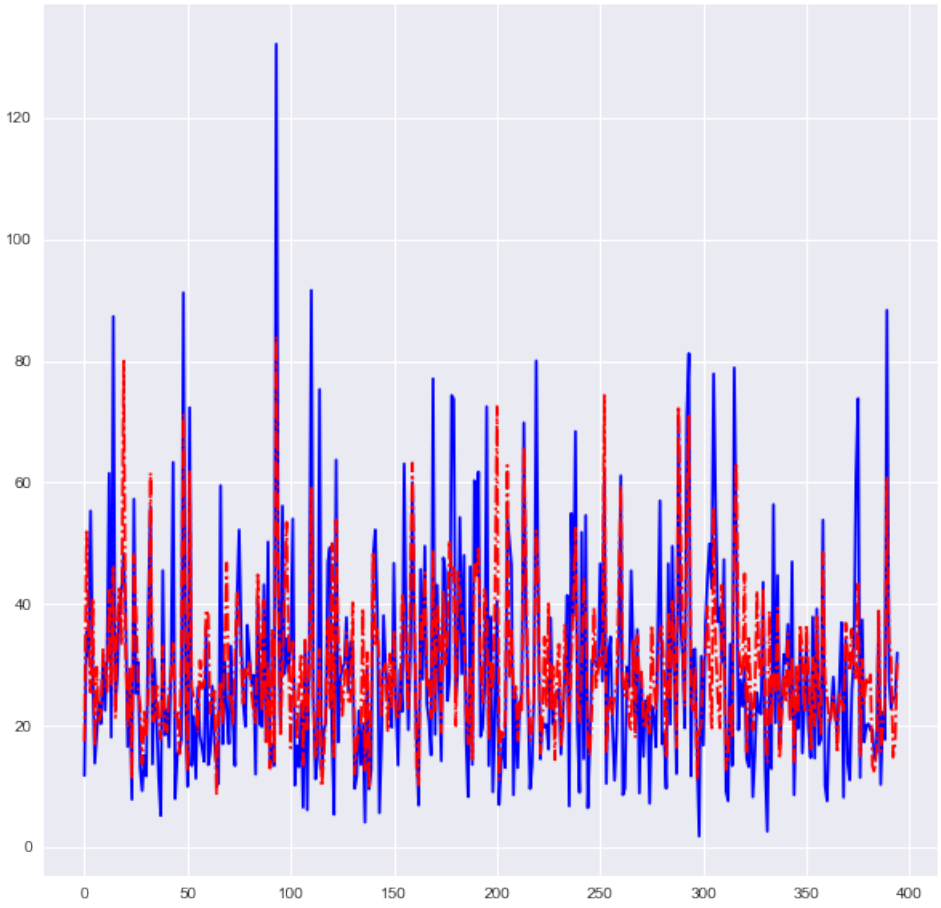
Zmienne pogodowe z logarytmowaniem:

R2 na zb. testowym: 0.497 Kor. Spearmana na zb. testowym: 0.671 Kor. Pearsona na zb. testowym: 0.699

Zauważmy, że po raz kolejny logarytmowanie nie polepsza modelu, a tutaj wręcz go pogarsza. W dalszej części używane są tylko zmienne niezlogarytmowane.

Dokładając smog z dnia poprzedniego nieznacznie polepszylismy precyzję predykcji modelu.

Wykres modelu bez logarytmowania, widać lekką poprawę w stosunku do poprzedniego:



Rysunek 6.2: Pokrycie modelu bazującego na zmiennych pogodowych i poprzedniej wartości zapylenia

6.5 Ostateczne modele

Ostatnim krokiem w modelowaniu jest wykonanie modeli dla 5 wybranych zmiennych objaśnianych:

- PM2,5 średnie Wokalna
- PM10 o 8 Niepodległości
- PM2,5 o 8 Niepodległości
- PM10 o 8 Wokalna
- PM10 średnie Niepodległości

PM2,5 średnie Wokalna:
R2 na zb. testowym: 0.721
Kor. Spearmana na zb. testowym: 0.87
Kor. Pearsona na zb. testowym: 0.849
PM10 o 8 Niepodległości:
R2 na zb. testowym: 0.329
Kor. Spearmana na zb. testowym: 0.527
Kor. Pearsona na zb. testowym: 0.575
PM2,5 o 8 Niepodległości:
R2 na zb. testowym: 0.419
Kor. Spearmana na zb. testowym: 0.647
Kor. Pearsona na zb. testowym: 0.647
PM10 o 8 Wokalna:
R2 na zb. testowym: 0.312
Kor. Spearmana na zb. testowym: 0.603
Kor. Pearsona na zb. testowym: 0.569
PM10 średnie Niepodległości:
R2 na zb. testowym: 0.497
Kor. Spearmana na zb. testowym: 0.736
Kor. Pearsona na zb. testowym: 0.705
Zauważmy:

- PM2,5 jest przewidywane trochę lepiej niż PM10
- Smog uśredniony dla całego dnia przewidywany lepiej niż dla 8 rano
- Smog na Wokalnej przewidywany jest lepiej niż na Niepodległości

Rozdział 7

Badanie otrzymanych modeli

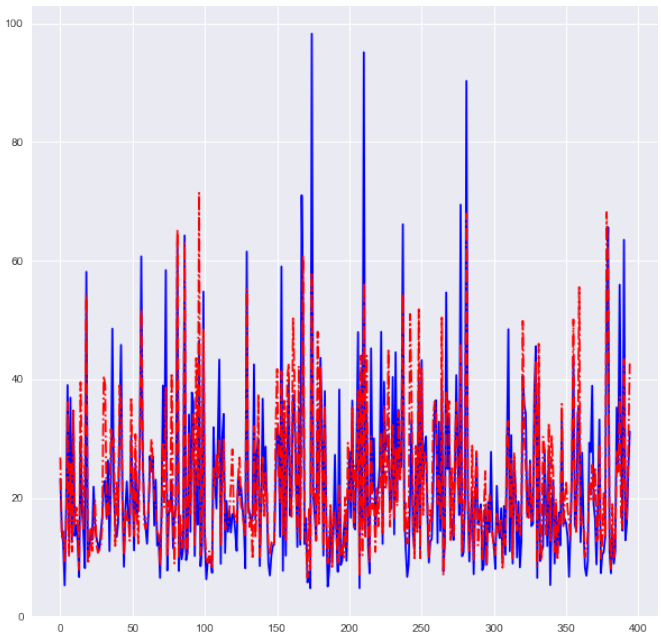
7.1 Przedstawienie najlepszego modelu

Model o największej precyzji predykcji to model dla PM2,5, smogu uśrednionego, mierzonego na Wokalnej.

R2 na zb. testowym: 0.721

Kor. Spearmana na zb. testowym: 0.87

Kor. Pearsona na zb. testowym: 0.849



Rysunek 7.1: Pokrycie najlepszego modelu

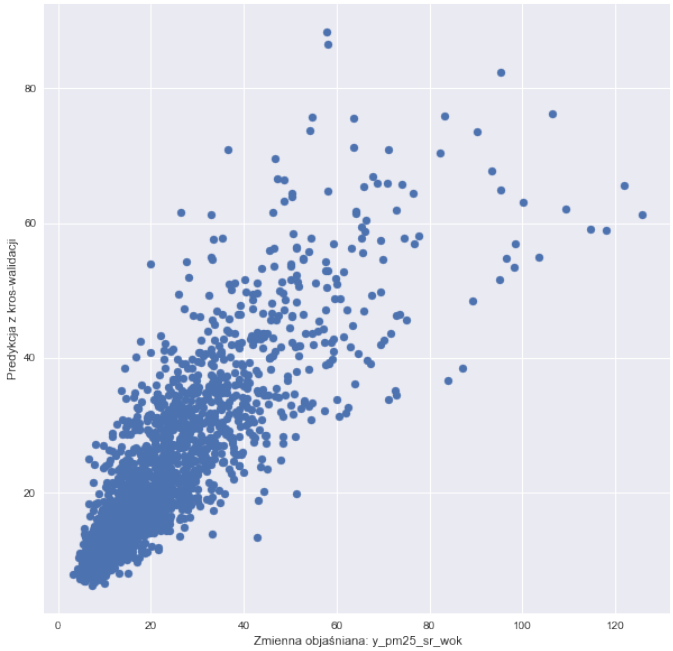
7.2 Kros-walidacja modelu

Testując stworzony model, autor przeprowadził testy kros-walidacyjne, czyli wtórny podział zbioru na treningowy i testowy i wywołanie modelu na nowych danych, sprawdzając jak zmienia się precyzja predykcji.

R2 dla zbiorów treningowych z kros-walidacji - testy przeprowadzone zostały 6 razy:

- 0.72253292
- 0.66410089
- 0.71541177
- 0.64480448
- 0.72832458
- 0.61368826

Model ma rzeczywistą predykcję na poziomie ok. 70%.



Rysunek 7.2: Predykcja z kros-walidacji - zmienna objaśniana

7.3 Interpretacja modelu predykcji

Głównym narzędziem interpretacyjnym stworzonego modelu są ważności cech które wchodzą do modelu:
[(0.51849999999999996, 'pop_pm25_sr_wok')
(0.1356, 'sr_predk_wiatr')
(0.059200000000000003, 'poprz_sr_temp')
(0.050700000000000002, 'sr_temp')
(0.042700000000000002, 'rozn_wilg')
(0.030499999999999999, 'sr_cisn_morze')
(0.028199999999999999, 'sr_zachm')
(0.026700000000000002, 'rozn_temp')
(0.024799999999999999, 'rozn_pr_wiatru')
(0.0224, 'rozn_cisn')
(0.0178, 'rozn_zachm')
(0.017000000000000001, 'sr_wilg')
(0.0106, 'rozn_opad')
(0.0085000000000000006, 'miesiac')
(0.0054999999999999997, 'suma_opad')
(0.0011999999999999999, 'bez_deszczu')
(0.0, 'bez_wiatru')]

7.4 Wnioski

Najważniejsze wnioski dotyczące procesu predykcji zapylenia wyciągnięte dzięki modelowaniu:

- Prędkość wiatru jest najważniejszym czynnikiem mającym wpływ na poprawną predykcję
- Temperatura w dniu analizowanym i w dniu poprzednim to drugi najważniejszy czynnik
- Zmiana wilgotności z dnia na dzień również w duży sposób wpływa na predykcję
- Ciśnienie atmosferyczne i zachmurzenie mają już mniejszy wpływ, ale też są istotne
- Możemy uznać, że pozostałe czynniki nie mają wpływu na predykcję smogu

- Łatwiejsze jest prognozowanie smogu w obszarze mieszkalnym (Wokalna na Ursynowie), gdzie ruch uliczny jest mniejszy, a palenie w piecach ma zdecydowanie największy wpływ na smog, niż w obszarze o dużym ruchu ulicznym (Niepodległości), ponieważ nie mam jak "uchwycić"tego czynnika w modelu

Rozdział 8

Podsumowanie

8.1 Wstęp do podsumowania

Przedmiotem pracy było stworzenie modelu wyjaśniającego jak czynniki pogodowe wpływają na poziom zapylenia smogowego. Stworzony model ma służyć predykcji smogu w analizowanym dniu w oparciu o dane z przeszłości.

Podczas procesu tworzenia pracy, zostały wykonane następujące etapy:

1. Zebranie danych historycznych dotyczących pogody i zapylenia powietrza.
2. Obróbka danych do postaci pozwalającej na analizę.
3. Eksploracyjna analiza danych i znalezienie relacji pomiędzy cechami(zmiennymi).
4. Zbudowanie modelu predykcyjnego z użyciem uczenia maszynowego.
5. Testowanie modelu i interpretacja wyników

8.2 Podsumowanie

Wszystkie założenia pracy opisane w poprzedniej sekcji zostały spełnione. Stworzony model w stopniu zadawalającym rozwiązuje problem predykcji dla konkretnego przypadku: obszary typowo mieszkalne na obszarach zurbanizowanych, bazując na smogu uśrednionym z całej doby.

Kolejnymi krokami mogłoby być stworzenie aplikacji pobierającej prognozę pogody na następny dzień i wywoływanie modelu na takich danych, sprawdzając później czy prognoza się sprawdziła i aktualizować model w oparciu o to.

Sam model poprawić możnaby, biorąc pod uwagę również dane z ruchu ulicznego dla różnych części miasta. Uśrednianie smogu dla całego dnia również okazało się sporą przeszkodą, użycie danych godzinowych dla pogody i analiza dobowa smogu mogłaby dać lepszą predykcję.

Bibliografia

[1] http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/

[2] https://www.eea.europa.eu/media/newsreleases/many-europeans-still-exposed-to-air-pollution-2015/premature-deaths-attributable-to-air-pollution

[3] Air pollution prediction via multi-label classification; G. Corani, M. Scanagatta; https://www.sciencedirect.com/science/article/pii/S1364815216300500?via%3Dihub

[4] Data mining methods for prediction of air pollution; K. Siwek, S. Osowski; http://pldml.icm.edu.pl/pldml/element/bwmeta1.element.bwnjournal-article-amcv26i2p467bwm