Разведочный анализ датасетов по шести загрязнителям, предоставленных Нидерландами, за период с 2013 по 2022 годы

1. jupiter-файлы EDA

jupiter-файл EDA с таблицами, графиками и выводами можно найти по следующе ссылке:

https://github.com/AlexandraNasonova/air_pollution_predict/blob/master/airpollpredictor/eda/aqeda.ipynb

Выводы в данном документе по больше части повторяют выводы в jupiter-файле.

2. Анализ пропущенных, невалидных, повторяющихся значений

2.1. В датасете не всегда указаны значения концентраций (столбец Concentration). Их доля зависит от загрязнителя и составляет от 3 до 7%.
Строки с неуказанными концентрациями необходимо оставить для сохранения последовательности непрерывной.

2.2. Невалидированные данные (Validity <= 0 или не определено), для которых концентрация указана, в датасете не найдены.
В соответствии с документацией к датасету, невалидными являются данные за пределами измерений приборов и проч. В данной работы не стоит цель оценки качества работы приборов, поэтому если такие данные будут найдены в датасетах по другим старанам, то концентрации необходимо будет заменить на NaN и рассматривать впоследствии как отсутствующие.

2.3. Есть неверифицированные или непрошедшие полный процесс верификации данные (Verification != 1). Их доля зависит от загрязнителя и составляет от 0 (для NO2) до 24% (для PM2.5). Такие данные не должны входить в формальные отчеты. Но в нашей учебной работе исключать их необходимости нет. Поэтому мы их будем рассматривать как верифицированные данные.

2.4. Минимальные значения Concentration меньше нуля или равны нулю. Такие концентрации не имеет физического смысла, но, в соответствии с изученными регламентами, можно утверждать, что в датасетах допускается небольшое количество отрицательных значений (если так отработал датчик, а показания должны быть сданы).
Доля отрицательных значений в основном не от 0 до 2,7 %, кроме SO2 с 9%.
Доля нулевый значений в основном не более 1%, кроме SO2 с 3%.

Построены распредления отрицательных значений. Пики этих распределений находятся вблизи нуля, но встречаются значения существенно меньше нуля. Последние значения принято решение считать невалидными.
Таким образом, значения в близи нуля требуется привести к нулю, считая приборной погрешностью. Значения существенно выше нуля, можно счесть невалидными и проставить для них NaN.

Что касается SO2, далее в ходе анализа был построен временной ряд по одной из станций измерения SO2. И на этом временном ряду видно, что до 2018 года эта станция выдавала очень много отрицательных концентраций. После 2018 года временной ряд стал совершенно другим, без заметных на графике отрицательных концентраций. Видимо поменяли оборудование или еще каким-то образом повлияли на качество измрений. Поэтому, возможно имеет смысл исключить данную станцию из прогноза.

2.5. Были найдены 2 строки с неуказанными значениями для кода станции и даты измерения. Эти строки были удалены, их никак невозможно учесть в прогнозе.

2.6. Для всех загрязнителей, кроме PM2.5 и PM10, имеется только один тип временного интервала измерения концентраций - по часам (столбец AveragingTime). В случае с PM2.5 и PM10 вторым интервалом является день, других интервалов нет.
С целью сохранения временной последовательности непрерывной и однотипной, было принято решение дополнить датасет строками на каждый час на день до DatetimeEnd строки с интервалом day. В новые строки записать концентрации NaN, а другие значения взять из строки с интервалом day.

2.7. Было обнаружено, что одна станция на один час может выдавать 2 значения Концентрации. Эти значения могут быть как одинаковыми, так и различными. Доля таких значений доходит до 30% в случае загрязнителя NO2. Анализ показал, что более 2х значений не бывает. Дело в том, что одной станции может соответствовать 2 различных значения SamplingProcess. Поскольку эти данные не являются полными дубликатами и несут полезную информацию, то решено их оставить.

3. Анализ выбросов

3.1. Анализ выбросов был проведен в по каждой станции в отдельности. На boxplot-диаграмах видно, что есть станции на которых выбросы существенны, на 1-2 порядка превышают средние значения.

3.2. Доля выборосов в разрезе загрязнителей составляет от 1 до 7% (SO2).

3.3. Есть станции с выбросами значительно большими, чем на других станциях. Причем и количество таких выбросов не единично, что может свидетельствовать о том, что это часть их технологического процесса, а не случайность. В таком случае, отбрасывать их нельзя, возможно удастся выделить сезонность.

3.4. Наиболее значительны выбросы у PM2.5, PM10, SO2 (на 2 порядка). Данные загрязнители являются продуктом работы индустриальных предприятий и вполне могут соответствовать реальным значениям, связанным с нештатными ситуациями. Их необходимо учитывать и рассмотреть отдельно.

3.5. Как правило, концентрации в разрезе станций имеют положительно скошенное распределение с одним пиком. Значения среднего и медианы близки, медиана немного меньше.

3.6. У O3 имеется второй пик вблизи нуля.

4. Анализ корреляций между загрязнителями

4.1. Видна очень сильная корреляция между PM2.5 и PM10, что понятно, потому что у них один источник - сжигание угля и проч. К тому же возможно при измерении PM10 частично учитываются и PM2.5.

4.2. Есть умеренная корреляция между CO и PM2.5/PM10, , что понятно, потому что у них один источник - сжигание угля и проч.

4.3. Корреляция между SO2 и PM2.5/PM10 очень слабая, хотя у них предположительно также один источник.

4.4. Видна сильная концентрация между CO и NO2, что понятно, потому что у них один источник - сжигание дизельного топлива.

4.5. Имеется довольная сильная отрицательная корреляция между O3 и NO2. Связано это с механизмом формирования ground level ozone (O3) из (в том числе) NOx. Таким образом O3 преобразовываясь из NO2 снижает концентрацию последнего.

5. Анализ в разрезе метаданных по станциям

5.1. Из метаданных по станциям были взяты 2 классификатора. Превый - это тип местности, где размещены станции: rural (сельскохозяйственная) и urban (урбанизированная), suburban (промежуточная). Второй - направленность станции: background (нет доминирующих источников), industrial (источники - заводы, теплостанции и проч) и traffic (источники – транспорт).

5.2. Анализ концентраций в разрезе типа местности.

5.2.1. В среднем концентрации O3, PM2.5, PM10 не зависят от типа местности.

5.2.2. Концентрации CO в урбанизированной местности немного выше, чем в сельской. По suburban местности данных нет.

5.2.3. Концентрации SO2 максимальна в урбанизированной местности (в 2 раза выше, чем в сельской), что вероятно связано с наличием там заводов и тепловых станций. Почему в suburban местности средняя концентрация почти в 2 раза меньше, чем в сельской, неясно.

5.2.4. Концентрации NO2 максимальна в урбанизированной местности (в 1.5 раза выше, чем в сельской), и немного выше в suburban. Что вероятно связано с большим количеством транспорта в них.

5.3. Анализ концентраций в разрезе направленности станций.

5.3.1. Концентрации O3 для станций background несколько выше, но существенно для разных направленностей станций не отличаются.

5.3.2. Концентрации PM2.5, PM10, CO существенно для разных направленностей станций не отличаются, в industrial / traffic немного выше.

5.3.3. Данных о SO2 для станций traffic нет. Для industrial в 3 раза выше, чем для и background.

5.3.4. Концентрации NO2 в среднем почти в 2 раза выше для станций traffic, чем для background, и почти в 1,5 раза выше для станций industrial, чем для background

6. Временные ряды концентраций загрязнителей

6.1. На графике концентрации O3 видна сезонность с пиками летом и минимума зимой. Вероятно это связано с механизмом образования O3 через NOx соединения, которые могут выделяться при аграрных и вегетационных процессах. Преимущественно эти процессы приходятся на лето и высокие температуры. Также на лето приходятся выбросы, причем сравнивая 2 станции складывается впечатление, что выбросы приходятся на одни и те же даты. Возможно жо связано с какими-то массовыми сельхоз работами.

6.2. На графике концентрации PM2.5 также видна сезонность, но видна не так явно, требуется проверка. Небольшой подъем приходится на начало года. Поскольку PM2.5 активно выделяется при сжигании топлива, то повышенные концентрации могут быть связаны с активным отопительным сезоном, т.е. с периодами низких температур. Так же четко на начало каждого года приходятся огромные выбросы. Есть предположение, что это связано с рождественскими фейерверками.

6.3. График концентрации PM10 очень похож на график PM2.5, что и понятно, потому что эти загрязнители производятся одними и теми же технологическими и природными процессами. Но для PM10 есть большие выбросы на начало года, но есть и другие. Надо посмотреть на какие даты они приходятся. Так же приведен нетипичный график, станция находится на границе города по тем же координатам, что и станция измерения PM2.5 с графика с "рождественскими" выбросами, что неожиданно. Может быть оборудование настроено иначе.

6.4. График концентрации CO также на первый взгляд имеет сезонность с подъемами зимой и спадами летом. Вероятно это связано с более активным сжиганием топлива в отопительные сезоны.

6.5. График концентрации SO2 возможно имеет убывающий тренд. Кажется также, что есть некоторая сезонность с убыванием концентрации летом. На одном из графиков видно, что внезапно изменилась структура данных, стало меньше отрицательных значений. Возможно сменили оборудование.

6.6. На график концентрации NO2 видно, что выбросы невелики. Это можно объяснить тем, что основным источником NO2 являются дизельные двигатели и ситуаций, когда их начинают внезапно использовать намного больше обычного, нетипичны. Возможно есть сезонность с подъемом зимой и спуском летом, что можно также объяснить особенностями работы двигателей при низких температурах. Данных после 2018 года нет, объяснения этому пока не найдены.