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

Для Нидерландов (NL) EDA выполнила Александра Насонова

Для Дании (DK) и Сербии (RS) EDA выполнила Адель Ким (Мансур Шамсутдинов)

Выводы по каждой работе были сгруппированы в соответствующие подпункты с названиями, соответствующими кодам стран.

1. Jupyter-файлы EDA

NL:

Jupyter-файл EDA с таблицами, графиками и выводами можно найти по следующей ссылке:

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

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

DK+RS:

Jupyter-файл EDA по Сербии и Дании можно найти по следующей ссылке:

https://github.com/AlexandraNasonova/air_pollution_predict/blob/master/aqi_eda/aqi_eda_03_analysis_RS_DK.ipynb

2. Анализ пропущенных, невалидных, повторяющихся значений

NL:

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 странам, то концентрации необходимо будет заменить на NaN и рассматривать впоследствии как отсутствующие.

2.3. Есть неверифицированные или не прошедшие 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 нулевых значений в основном не более 1%, кроме SO2 с 3%.

Построены распределения отрицательных значений. Пики этих распределений находятся вблизи нуля, но встречаются значения существенно меньше нуля. Последние значения принято решение считать невалидными.
Таким образом, значения вблизи 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 измерений. 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.

DK+RS:

2.8. Первичный анализ данных по совместному датасету Дании и Сербии показал, что в данных по концентрациям есть пропуски, отрицательные значения и очень большие по модулю значения. Также что для всех загрязнителей больше всего данных приходится на станции расположенные в городах (urban в колонке «AirQualityStationArea»). Также оказалось, что в датасете есть по разному агрегированные данные («AveragingTime») - hour, day, var.

2.9. В ходе дальнейшего анализа выяснилось, что в датасете есть невалидные данные, причем обнаружилась прямая корреляция пропусков и невалидных данных. Доля таких данных составляла примерно 6% от общего датасета. Невалидные данные были удалены из датасета, что одновременно избавило нас и от пропусков. Также из датасета были исключены все данные с «AveragingTime» day и var. В нашей работе планируется использовать только данные с «AveragingTime» hour. Данные агрегированные в другом временном интервале встречаются

не во всех станциях и не имеют какой-либо ценности для нас, так как они рассчитаны на основе имеющихся данных hour.

2.10. В датасете есть неверифицированные данные. Встречаются они повсеместно и доля таких данных высока, особенно в данных за последние несколько лет, поэтому исключить такие данные мы не можем. Оценка характеристик верифицированных и неверифицированных данных не выявила каких-либо существенных отличий в плане качества данных.

2.11. Оценка долей отрицательных значений показала что на Данию приходится существенно больше таких значений чем на Сербию. После анализа отрицательных значений выяснилось, что основная их часть располагается в довольно узком диапазоне недалеко от нуля, но среди них, также как и с положительными данными, встречаются выбросы, расположенные сильно далеко от основной группы данных. Изначально рассматривалось три разных способа устранения проблемы с отрицательными данными:
1) Удалить такие данные, или удалить станции в которых они встречаются
2) Заменить их на модуль их значения
3) Заменить их на ноль или близкое к нулю значение.
Каждый вышеупомянутый вариант имеет свои недостатки. Так например удалив такие данные мы бы создали пропуски в timeseries, которые пришлось бы как-то восстанавливать. Вариант удалить конкретные станции оказался не очень применим, т.к такие данные встречаются в большинстве станций. Другим возможным решением могло бы быть замещение отрицательных значений их модулями. В данной работе даже было проверено, что для большинства отрицательных значений их модули не превосходили верхние отсекающие границы (усы боксплота) общих выборок каждого загрязнителя. А учитывая их небольшую долю относительно основной выборки, мы бы, наверное, не сильно изменили характеристики основной выборки.
После чтения доступной информации по особенностям работы оборудования измерительных станций выяснилось, что такой способ исправить отрицательные данные был бы скорее всего не совсем корректным. Оказалось, что отрицательные значения могут возникать в связи с плохо откалиброванными детекторами, а также в тех случаях, когда концентрация измеряемого загрязнителя меньше или близка к нижнему пределу обнаружения прибора. Большое количество отрицательных значений в течение продолжительного времени может свидетельствовать о систематической проблеме с качеством калибровки оборудования данной станции. А это уже, в свою очередь, ставит под сомнение качество и валидность в том числе и положительных измерений. Т.к. в нашем датасете для Дании было существенно больше таких измерений, то это стало одной из причин по которой было решено исключить все данные по Дании и продолжить дальнейший анализ на данных по Сербии.

2.12. В данных по Дании для некоторых станций были обнаружены дубликаты измерений с одинаковым timestamp'ом. Для Сербии дубликаты не выявились. Однако после чтения информации об особенностях работы измерительных станций, выяснилось, что они могут иметь по два детектора для загрязнителей. И возможно такое задвоение связано с тем, что экспортированные данные с таких станций содержали измерения обоих детекторов по отдельности. В данной работе не были проанализированы колонки "Samplingpoint", "SamplingProcess" и "Sample". Возможно, после анализа данных в этих колонках можно было бы найти оптимальное решение по обработке таких "дубликатов". Дальнейший анализ было решено провести на данных Сербии, для которых проблема с дубликатами отсутствовала.

3. Анализ выбросов

NL:

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

3.2. Доля выбросов 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.

RS:

3.7. Была рассчитана доля выбросов каждого загрязнителя в датасете по Сербии. Для большинства загрязнителей доли выбросов были относительно невысокими (0.35%-8.7%). Однако для SO2 и CO, рассчитанные доли выбросов оказались равными 23% и 74% соответственно. Сербия была выбрана нами как пример страны с низким качеством воздуха, и вполне возможно, что полученные доли не являются аномалиями а реально связаны с периодическими выбросами высоких концентраций загрязнителей в воздух. В дальнейшем планируется оценить доли загрязнителей в разрезе типов станций, в разрезе их расположения, а также провести индивидуальный анализ отдельных станций. Вполне возможно, что несколько «аномальных» станций, расположенных в индустриальных зонах искажают общую картину по выбросам.

4. Анализ корреляций между загрязнителями

NL:

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.

RS:

4.6. В ходе данной работы изначально не планировалось проводить анализ корреляций между загрязнителями. Так как маловероятно что они существенным образом будут отличаться от общеизвестных корреляций данных загрязнителей описанных в литературе и маловероятно что они существенным образом будут отличаться от результатов, полученных на данных Нидерландов. Но чуть позже такой анализ будет проведен для 1 выбранной станции Сербии и соответствующие выводы будут добавлены в данный раздел.

5. Анализ в разрезе метаданных по станциям

NL:

5.1. Из метаданных по станциям были взяты 2 классификатора. Первый - 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

RS:

5.4. Для SO2 каких-либо существенных отличий как в разрезе типов местности, так и в разрезе типов станций выявлено не было.

5.5. Для PM10 в категории suburban медианное значение концентрации выше, чем в urban. В разрезе типов станций заметно выделяется категория industrial. traffic и background имеют примерно одинаковые медианные значения.

5.6. Для O3 наблюдается очень сильный перекос в сторону rural относительно urban (больше чем в два раза). При этом данные для категории suburban отсутствуют. Возможно это связано с тем, что в Сербии большинство предприятий, вырабатывающих данный загрязнитель располагается в удаленных от городов местах. При этом в разбивке по типам станции отсутствует тип industrial, а медианное значение background станций почти в 5 раз выше чем в traffic.

5.7. Для NO2 наибольшая медианная концентрация наблюдается в urban, она больше чем в два раза выше чем в suburban. А suburban, в свою очередь, больше чем в два раза чем в rural. Т.е. прослеживается довольно четкая зависимость медианной концентрации от удаленности станций от городов. Скорее всего это связано с тем, что данный загрязнитель вырабатывается преимущественно при сжигании автомобильного топлива. В разрезе типов станций это косвенно подтверждается тем, что медианное значение traffic почти в два раза выше чем у категорий industrial и background (у которых медианные значение практически идентичны).

5.8. Для CO наблюдается незначительное поэтапное возрастание медианных значений при движении от категории rural к suburban и к urban. А при просмотре категорий видно, что в категории industrial медианное значение примерно на 20-25% ниже чем в других двух категориях(у которых они практически равны).

5.8. Для PM2.5 в категории suburban медианное значение концентрации выше, чем в urban, что хорошо коррелирует с результатами, полученными на PM10. Аналогично и в разрезе типов станций, точно так же выделяется категория industrial относительно traffic и background, которые имеют близкие медианные значения. Налицо явная и наглядная корреляция PM2.5 и PM10, что в принципе логично, так как по данным литературы данные загрязнители образуются при одних и тех же процессах.

6. Временные ряды концентраций загрязнителей

NL:

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 это связано с какими-то массовыми сельскохозяйственными 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.

RS:

6.7. В ходе данной работы изначально не планировалось проводить анализ временных рядов (так как он был проведен Александрой). Но так как нашей команде стало интересно насколько будут различаться временные ряды в случае выбора другой страны, то в ближайшее время будут построены timeseries по Сербии и выводы по результатам их оценки будут добавлены сюда.