Прогнозирование уровня загрязнения воздуха при помощи машинного обучения

Суслов Александр, к.т.н. Михаил Криницкий,
Prof. Chantal Staquet,Ph.D. Enzo Le Boudec

2024

План

Мотивация

Постановка задачи

Модели машинного обучения

Регрессия

Результаты регрессии

Классификация

Результаты классификации

Заключение

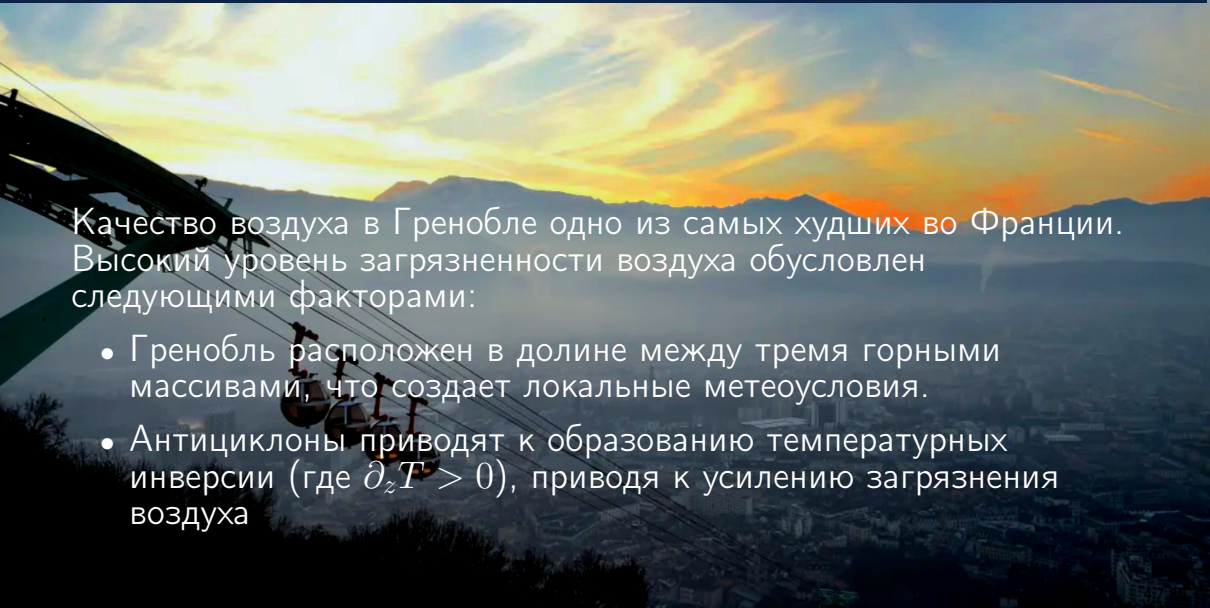
Мотивация

Загрязнение воздуха - проблема для большинства городов. Мелкодисперсные частицы (PM) является одним из основных загрязняющих факторов. Это смесь твердых и жидких частиц, находящихся в воздухе во взвешенном состоянии. Они делятся на 2 группы по размеру:

- Частицы диаметром менее 10 µm (particulate matter of diameters less than 10 micrometers) (PM10)
- Частицы диаметром менее 2.5 µm (PM2.5)

PM10 и PM2.5 оказывают негативное влияние на здоровье человека. Основные источники PM:

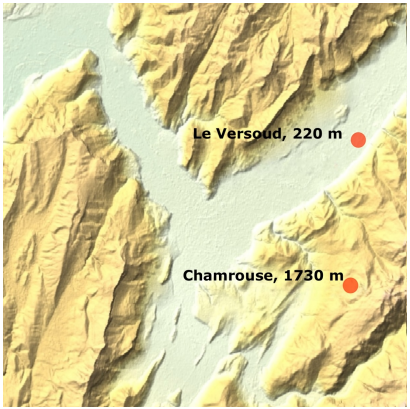
- Отопление
- Дорожное движение
- Производство

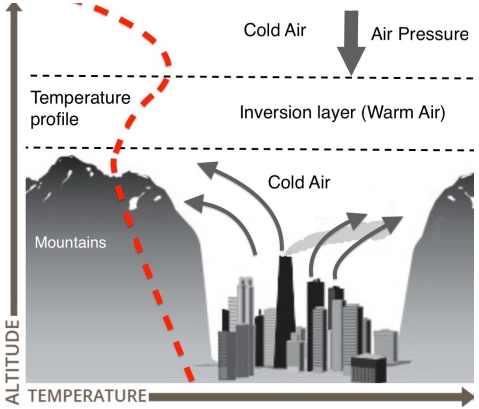


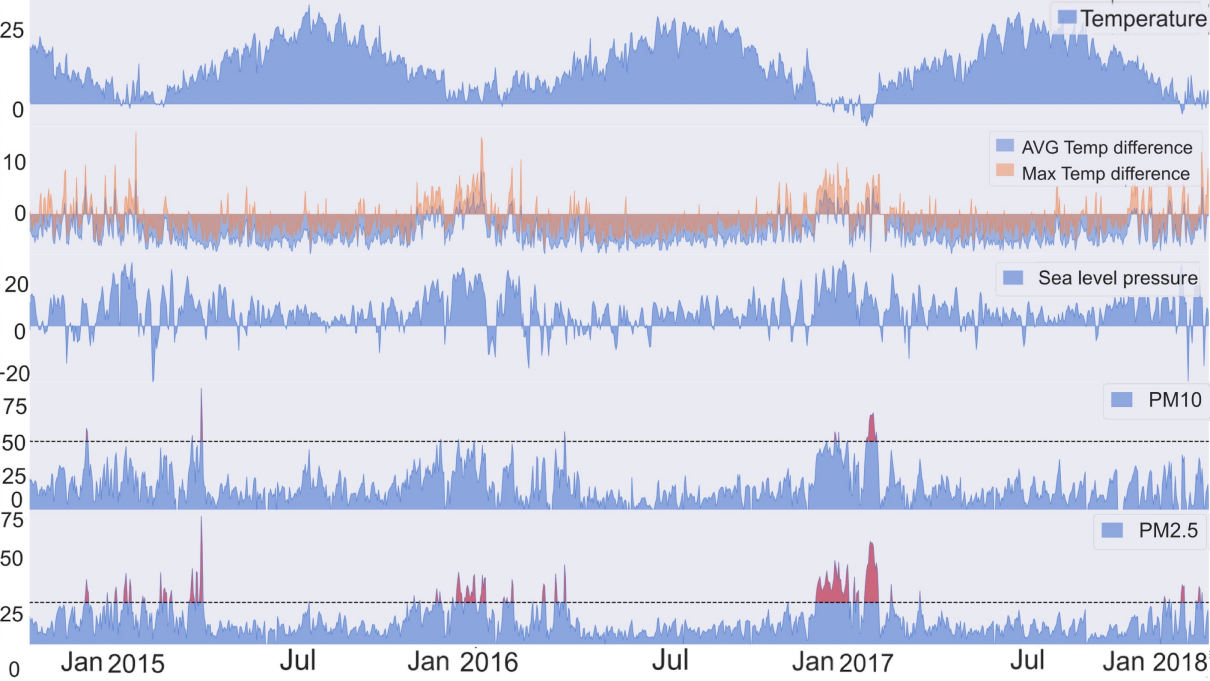
Качество воздуха в Гренобле одно из самых худших во Франции. Высокий уровень загрязненности воздуха обусловлен следующими факторами:

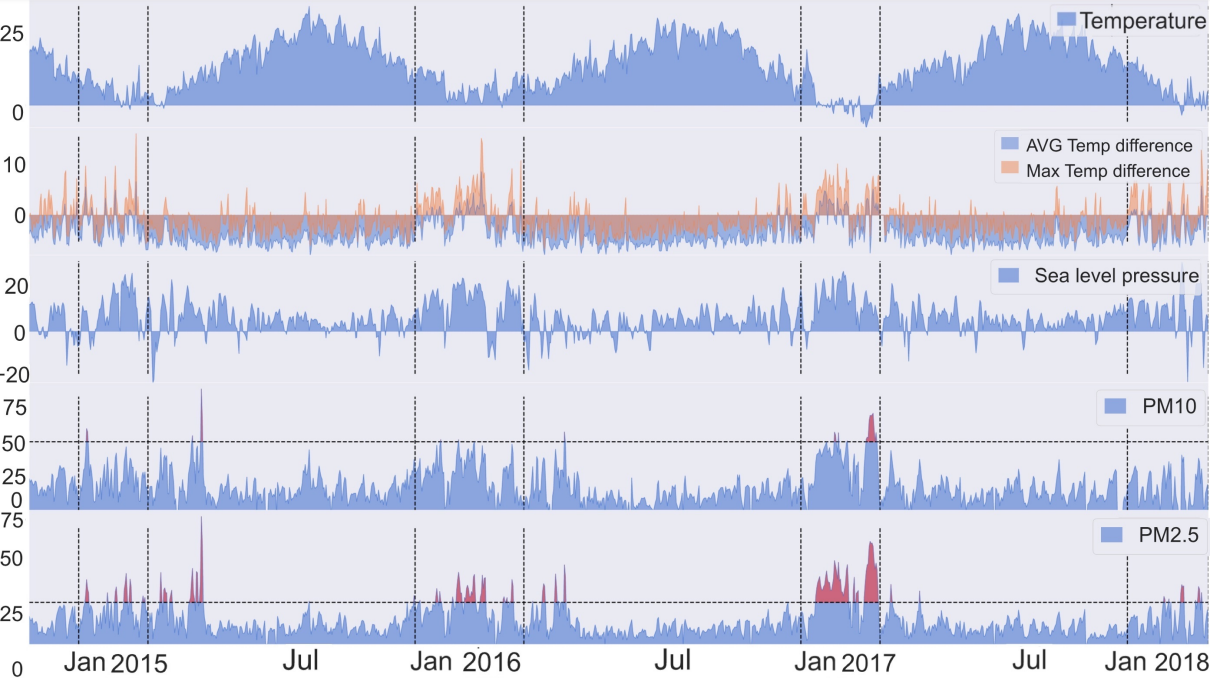
- Гренобль расположен в долине между тремя горными массивами, что создает локальные метеоусловия.
- Антициклоны приводят к образованию температурных инверсии (где $\partial_z T > 0$), приводя к усилению загрязнения воздуха

Для учета **температурной инверсии** мы рассмотрим разность температур между двумя станциями: Chamrousse (выс. 1730м) и Le Versoud (выс. 220м). Разность температур > 0 обусловливает инверсию.






Temperature
25
0
AVG Temp difference
Max Temp difference
10
0
Sea level pressure
20
0
-20
PM10
75
50
25
0
PM2.5
75
50
25
0
Jan 2015
Jul
Jan 2016
Jul
Jan 2017
Jul
Jan 2018


Temperature
25
0
AVG Temp difference
Max Temp difference
10
0
Sea level pressure
20
0
-20
PM10
75
50
25
0
PM2.5
75
50
25
0
Jan 2015
Jul
Jan 2016
Jul
Jan 2017
Jul
Jan 2018

Входные данные

Дневной порог концентрации PM согласно **ВОЗ**:

Pollutant	Daily Threshold
PM10	$50\,\mu g\,m^{-3}$
PM2.5	$25\,\mu g\,m^{-3}$

Цель: Прогноз концентрации PM на 3 дня вперед на основе фактических данных за последние 3 дня.

Входные данные:

- PM10 и PM2.5
- Скорость и направление ветра
- Осадки
- Температура
- Разность температур

Входные данные

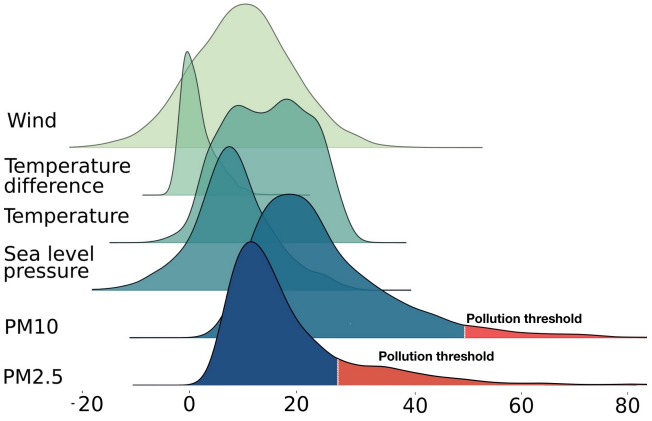
Дневной порог концентрации PM согласно **ВОЗ**:

Pollutant	Daily Threshold
PM10	50 $\mu g\, m^{-3}$
PM2.5	25 $\mu g\, m^{-3}$

Цель: Прогноз концентрации PM на 3 дня вперед на основе фактических данных за последние 3 дня.

Входные данные:

- PM10 и PM2.5
- Скорость и направление ветра
- Осадки
- Температура
- Разность температур



Входные данные

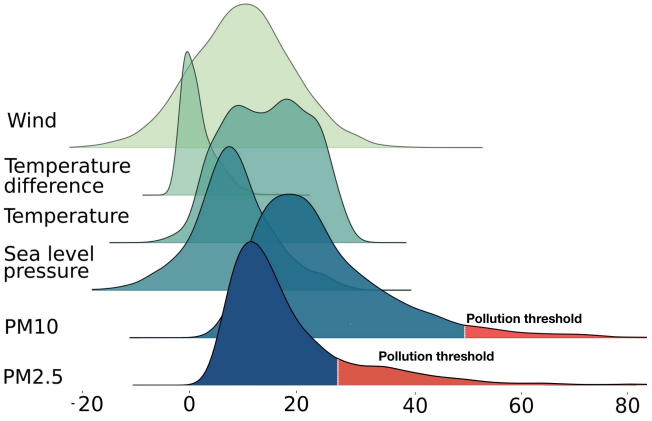
Дневной порог концентрации PM согласно **ВОЗ**:

Pollutant	Daily Threshold
PM10	$50\,\mu g\,m^{-3}$
PM2.5	$25\,\mu g\,m^{-3}$

Цель: Прогноз концентрации PM на 3 дня вперед на основе фактических данных за последние 3 дня.

Входные данные:

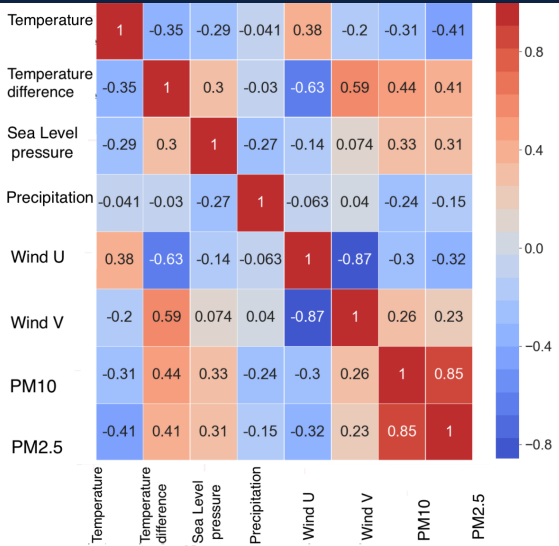
- PM10 и PM2.5
- Скорость и направление ветра
- Осадки
- Температура
- **Разность температур**



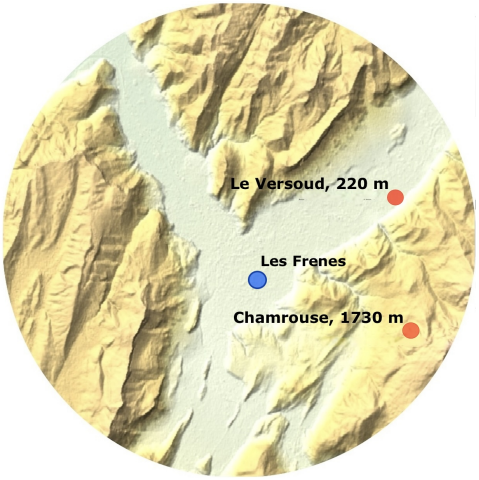
Входные данные

Для улучшения прогноза мы добавляем разность переменных между последовательными днями:

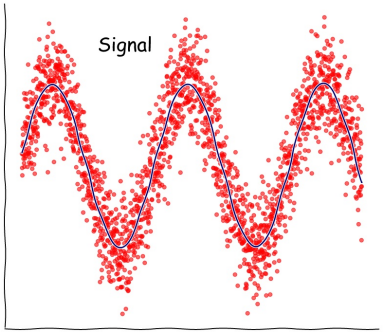
- ΔPM10 и $\Delta PM2.5$
- ΔСкорость и направление ветра
- Δ Давление
- ΔТемпература
- ΔРазность температур



Рассмотрим значения PM на станции Les Frenes, которая расположенна в парке:



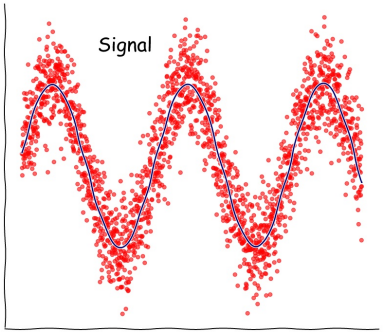
Аугментация данных



Проблема: Недостаточно данных в обучающей выборке.

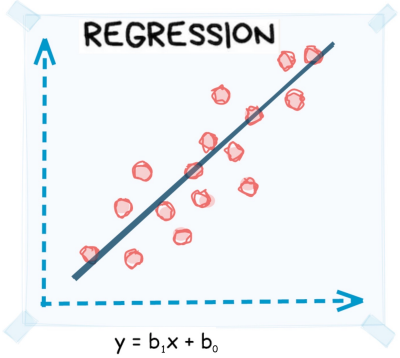
Решение: Аугментация обучающего набора данных искусственно сгенерированными данными с добавлением случайного шума. Особый акцент делается на загрязненных эпизодах.

Аугментация данных

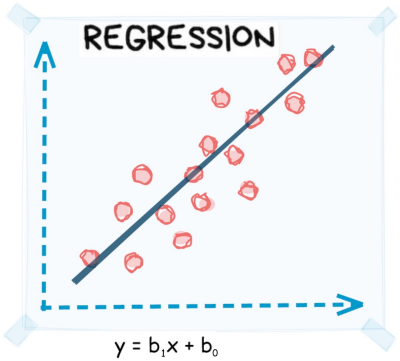


Проблема: Недостаточно данных в обучающей выборке.

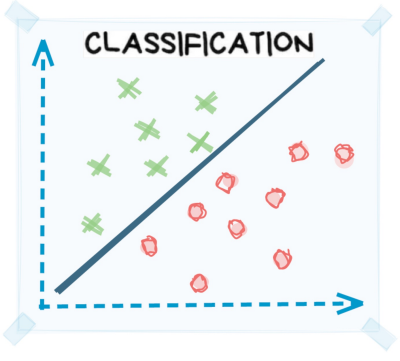
Решение: Аугментация обучающего набора данных искусственно сгенерированными данными с добавлением случайного шума. Особый акцент делается на загрязненных эпизодах.



Регрессия: Зависимая переменная непрерывная. Спрогнозировать значение концентрации PM на три дня вперед.



Регрессия: Зависимая переменная непрерывная. Спрогнозировать значение концентрации PM на три дня вперед.



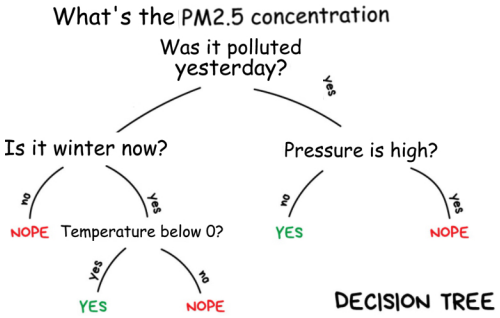
Классификация: Зависимая переменная-дискретная. Определить, превышают ли PM заданный порог на три дня вперед.

Модели

- **Базовые алгоритмы** используются для оценки эффективности более сложных моделей. Модель **Persistence** прогнозирует такой же уровень загрязнения, как и в предыдущий день.
- **Решающие деревья** разбивает данные, принимая решение, используя ряд вопросов. Каждый вопрос сужает подмножество возможных значений, пока модель не сумеет сделать прогноз.
- **Ансамбли моделей** объединяют предсказания нескольких моделей, которые по отдельности являются неточными, для получения более точного предсказания на новой выборке.

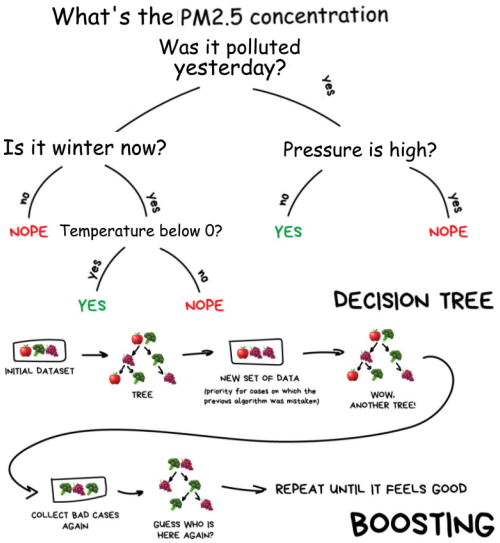
Модели

- **Базовые алгоритмы** используются для оценки эффективности более сложных моделей. Модель **Persistence** прогнозирует такой же уровень загрязнения, как и в предыдущий день.
- **Решающие деревья** разбивает данные, принимая решение, используя ряд вопросов. Каждый вопрос сужает подмножество возможных значений, пока модель не сумеет сделать прогноз.
- **Ансамбли моделей** объединяют предсказания нескольких моделей, которые по отдельности являются неточными, для получения более точного предсказания на новой выборке.



Модели

- **Базовые алгоритмы** используются для оценки эффективности более сложных моделей. Модель **Persistence** прогнозирует такой же уровень загрязнения, как и в предыдущий день.
- **Решающие деревья** разбивает данные, принимая решение, используя ряд вопросов. Каждый вопрос сужает подмножество возможных значений, пока модель не сумеет сделать прогноз.
- **Ансамбли моделей** объединяют предсказания нескольких моделей, которые по отдельности являются неточными, для получения более точного предсказания на новой выборке.



План

- Мотивация
- Постановка задачи
- Модели машинного обучения
- Результаты для регрессия
- Результаты для классификации
- Заключение

План

- Мотивация
- Постановка задачи
- Модели машинного обучения
- Результаты для регрессия
- Результаты для классификации
- Заключение

Регрессия: Прогнозирование непрерывного значения.

Задача: Прогнозирование точного значения концентрации PM на 3 дня вперед.

Метрики:

The Mean Absolute Error:

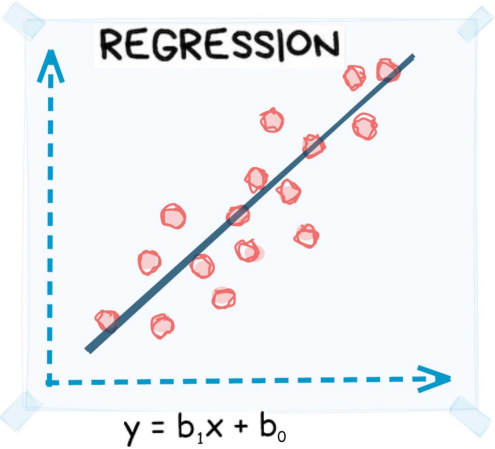
$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|P - M|$$

The Mean Squared Error:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(P - M)^2$$

The Root Mean Squared Error:

$$\mathrm{RMSE} = \sqrt{MSE}$$



Регрессия: Прогнозирование непрерывного значения.

Задача: Прогнозирование точного значения концентрации PM на 3 дня вперед.

Метрики:

The Mean Absolute Error:

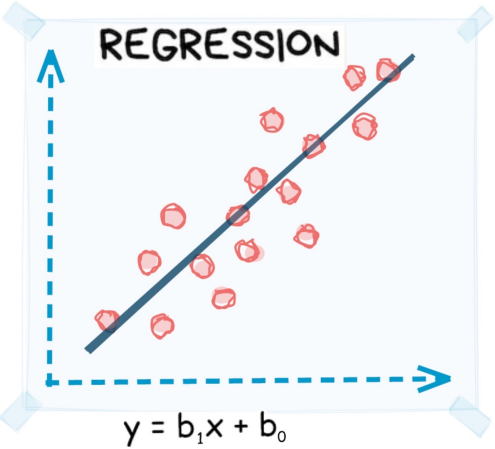
$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|P - M|$$

The Mean Squared Error:

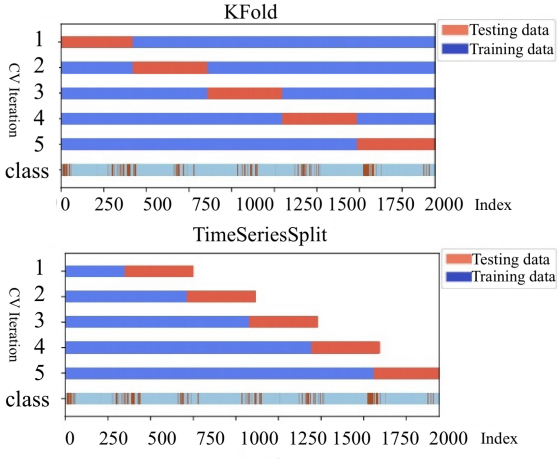
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(P - M)^2$$

The Root Mean Squared Error:

$$\mathrm{RMSE} = \sqrt{MSE}$$



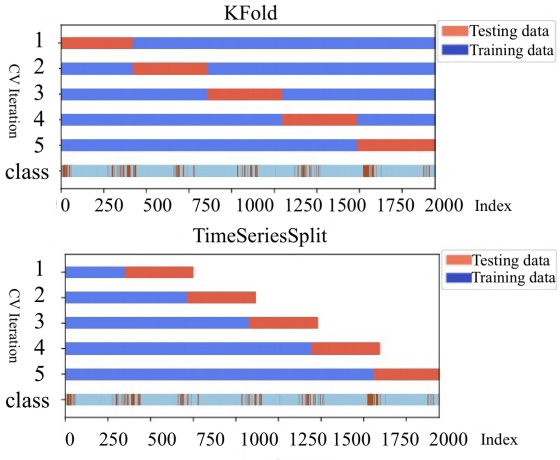
Оценка качества модели



Как надежно оценить качество модели?

Кросс-валидация разбивает данные на k частей. Одна из частей используется для тестирования, а другие k-1 - для тренировки модели. После тренировки и оценки модели этот процесс повторяется на каждой части тестовых данных.

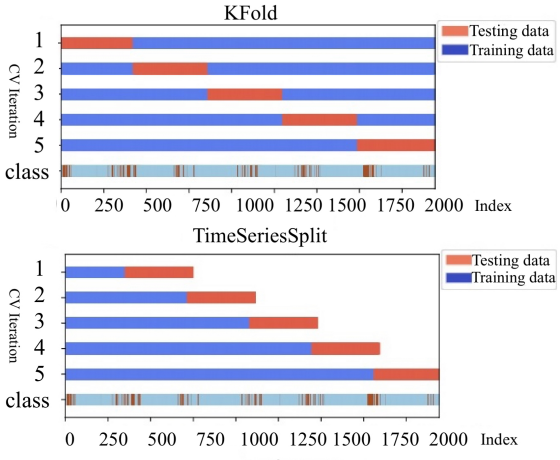
Оценка качества модели



Как надежно оценить качество модели?

Кросс-валидация разбивает данные на k частей. Одна из частей используется для тестирования, а другие k-1 - для тренировки модели. После тренировки и оценки модели этот процесс повторяется на каждой части тестовых данных.

Оценка качества модели



Как надежно оценить качество модели?

Кросс-валидация разбивает данные на k частей. Одна из частей используется для тестирования, а другие k-1 - для тренировки модели. После тренировки и оценки модели этот процесс повторяется на каждой части тестовых данных.

Аппроксимация приращений PM

Мы аппроксимируем разность PM между последовательными днями

Входные данные:

$\Delta PM_{day-1} = PM_{day-1} - PM_{day-2}$

$\Delta PM_{day0} = PM_{day0} - PM_{day-1}$

Зависимая переменная:

$\Delta PM_{day+1} = PM_{day+1} - PM_{day0}$

$\Delta PM_{day+2} = PM_{day+2} - PM_{day+1}$ и

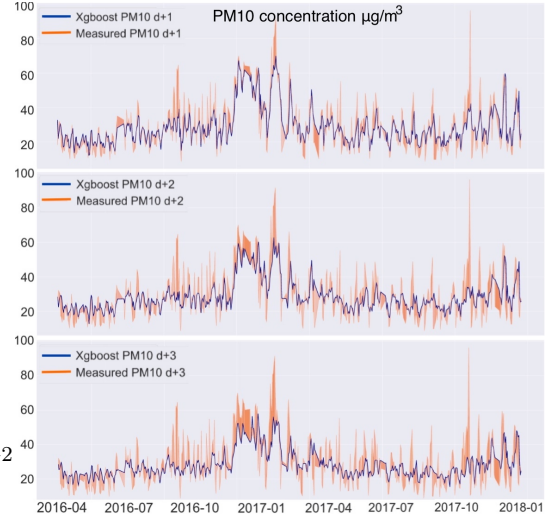
$\Delta PM_{day+3} = PM_{day+3} - PM_{day+2}$

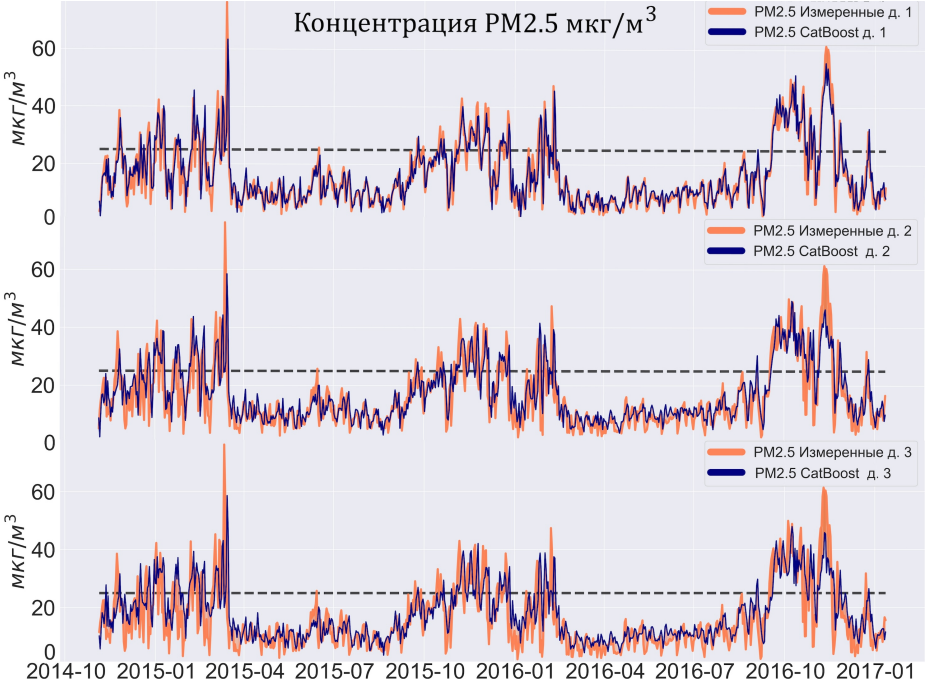
Исходные значения можно восстановить по формуле:

$PM_{day+1} = PM_{day0} + \Delta PM_{day+1}$;

$PM_{day+2} = PM_{day0} + \Delta PM_{day+1} + \Delta PM_{day+2}$

$PM_{day+3} = PM_{day0} + \Delta PM_{day+1} + \Delta PM_{day+2} + \Delta PM_{day+3}$




Концентрация PM2.5 мкг/м3
PM2.5 Измеренные д. 1
PM2.5 CatBoost д. 1
PM2.5 Измеренные д. 2
PM2.5 CatBoost д. 2
PM2.5 Измеренные д. 3
PM2.5 CatBoost д. 3
мкг/м3
0
20
40
60
2014-10
2015-01
2015-04
2015-07
2015-10
2016-01
2016-04
2016-07
2016-10
2017-01

Прогнозирование концентрации PM2.5 за весь период:

Results for PM2.5 at Les Frenes												
	day+1				day+2				day+3			
	MAE		RMSE		MAE		RMSE		MAE		RMSE	
Model	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
CatBoost	3.26	0.17	4.87	0.33	4.72	0.20	6.88	0.45	5.42	0.15	7.70	0.48
Linear regression	4.34	0.51	5.92	0.65	5.74	0.58	7.76	0.72	6.51	0.83	8.68	0.82
Persistence	3.90	0.41	5.71	0.62	5.60	0.69	8.00	0.98	6.35	0.83	9.01	1.16

Заключение Прогноз приращений PM между последовательными днями, вместо абсолютных значений приводит к улучшению качества прогноза для всех моделей, особенно в случае высокого уровня загрязнения. Тем не менее, с течением времени появляется определенный лаг между предсказаными и измеренными значениями.

Влияние разности температур

Использование разности температур в качестве признакового описания объекта позволяет учитывать температурную инверсию.

Разность между метриками при прогнозирования среднесуточного значения концентрации PM2.5 для зимнего периода с разностью температур в качестве входных данных модели и без:

day+1		day+2		day+3	
MAE	RMSE	MAE	RMSE	MAE	RMSE
0,82%	1,01%	1,0%	0,83%	0,97%	0,78%

Заключение Среднее улучшение метрик MAE и RMSE не превышает 1%. Возможная причина: другие переменные сильно коррелируют с разностью температур.

Постановка задачи

Классификация: Определение категории объекта.

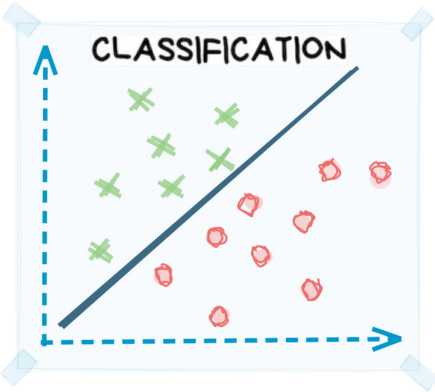
Задача: определить, будет ли воздух загрязнен в течение следующих трех дней. Дни с высоким уровнем загрязнения:

$$\langle PM10 \rangle_{24h} > 50\,\mu g\,m^{-3}$$
$$\langle PM2.5 \rangle_{24h} > 25\,\mu g\,m^{-3}$$

Несбалансированые данные.

Процент загрязненных дней для Les Frenes:

Station	Polluted days PM10	Polluted days PM2.5
Les Frenes	3.5%	15%



Постановка задачи

Классификация: Определение категории объекта.

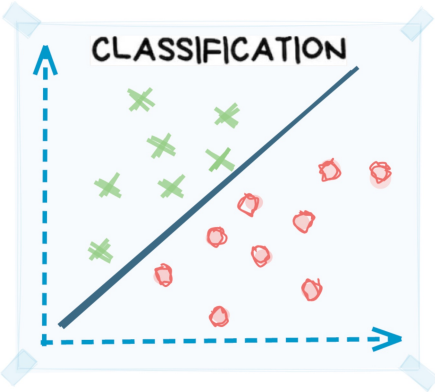
Задача: определить, будет ли воздух загрязнен в течение следующих трех дней. Дни с высоким уровнем загрязнения:

$$\langle PM10 \rangle_{24h} > 50\,\mu g\,m^{-3}$$
$$\langle PM2.5 \rangle_{24h} > 25\,\mu g\,m^{-3}$$

Несбалансированые данные.

Процент загрязненных дней для Les Frenes:

Station	Polluted days PM10	Polluted days PM2.5
Les Frenes	3.5%	15%



Accuracy - Доля правильных ответов от общего числа наблюдений.

Confusion matrix

		Predicted class	
Actual class		Class = No	Class = Yes
	Class = No	True negative	False positive
	Class =Yes	False negative	True positive

False Negative - Ошибка второго рода

False Positive - Ошибка первого рода

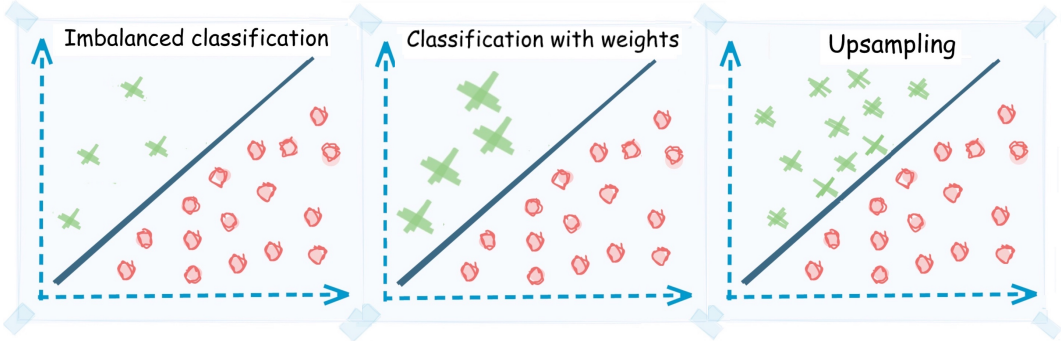
Precision - сколько из предсказанных положительных значений действительно положительные. $\mathrm{Precision} = \frac{\mathrm{T_p}}{\mathrm{T_p}+\mathrm{F_p}}$

Recall - доля корректно идентифицированных фактических положительных значений (True positive) $Recall = \frac{T_p}{T_p+F_n}$

F1-score - Среднее гармоническое Precision и Recall:

$\mathrm{F_1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$, где $F1 \in [0, 1]$

Классификация с весом



Вес - коэффициент, который вводится для того, чтобы модель отдавала предпочтение наименьшему классу. Статистически добавление веса эквивалентно искусственному дополнению выборки элементами из наименьшего класса. В нашем случае важнее определить загрязненные дни.

Мы преобразуем результат нашей регрессии следующим образом:

$$\text{Output}= \begin{cases} 1, \text{если } \langle PM2.5 \rangle_{24h} > 25\,\mu\text{g m}^{-3} \\ 0, \text{в остальных случаях} \end{cases}$$

F1-score для преобразованной регрессии:

PM2.5 results (whole year) at LF

	F1 day+1		F1 day+2		F1 day+3	
Model	Mean	Std	Mean	Std	Mean	Std
CatBoost	0.82	0.08	0.72	0.12	0.68	0.08
Persistence	0.74	0.08	0.62	0.11	0.54	0.16

Матрица ошибок для преобразованной регрессии:

Confusion matrix for PM2.5 (whole year) at LF

	Predicted					
Actual	d+1 False	d+1 True	d+2 False	d+2 True	d+3 False	d+3 True
False	1290	42	1264	69	1251	80
True	37	186	59	163	67	157

F1-score:

PM2.5 results (whole year) at LF						
	F1 day+1		F1 day+2		F1 day+3	
Model	Mean	Std	Mean	Std	Mean	Std
CatBoost	0.75	0.08	0.70	0.12	0.65	0.08
BRF	0.77	0.05	0.69	0.1	0.64	0.10
Persistence	0.74	0.08	0.62	0.11	0.54	0.16

Матрица ошибок для CatBoost:

Confusion matrix for PM2.5 (whole year) at LF						
	Predicted					
Actual	d+1 False	d+1 True	d+2 False	d+2 True	d+3 False	d+3 True
False	1342	64	1330	79	1336	74
True	48	**181**	60	**166**	75	**150**

Матрица ошибок для BRF:

Confusion matrix for PM2.5 (whole year) at LF						
	Predicted					
Actual	d+1 False	d+1 True	d+2 False	d+2 True	d+3 False	d+3 True
False	1248	158	1181	228	1104	306
True	23	**206**	29	**197**	19	**206**

Матрица ошибок для Persistence:

Confusion matrix for PM2.5 (whole year) at LF						
	Predicted					
Actual	d+1 False	d+1 True	d+2 False	d+2 True	d+3 False	d+3 True
False	1280	52	1255	78	1238	93
True	53	**170**	78	**144**	95	**129**

Заключение Если наша цель - правильно определить как можно больше эпизодов, то предпочтительнее использовать CatBoost. Если важнее не пропустить ни одного загрязненного эпизода, то предпочтительнее использовать BRF.

Station	Polluted days PM10	Polluted days PM2.5
Les Frenes	3.5%	**15%**

Средний уровень загрязнения и количество загрязненных эпизодов значительно выше в зимний период. Статистика по загрязнению по отопительным сезонам:

Season	Mean PM2.5	Max PM2.5	Days Total	Polluted days	Days under inversions	High atmos pressure days	Polluted days under inversion
Winter	20	76.25	1015	298	161	789	108
Summer	9.74	25.8	974	2	1	821	0

- 99% загрязненных дней приходятся на зиму.
- Более 30% из которых происходят при температурной инверсии.
- Средняя концентрация PM2.5 зимой на 100% выше.

Результаты для зимнего периода

F1-score для зимнего периода:

PM2.5 results (winter) at LF						
	F1 day+1		F1 day+2		F1 day+3	
Model	Mean	Std	Mean	Std	Mean	Std
CatBoost	0.74	0.08	0.66	0.14	0.62	0.12
BRF	0.76	0.06	0.65	0.13	0.60	0.13
Persistence	0.73	0.06	0.61	0.15	0.53	0.18

Матрица ошибок для CatBoost:

Confusion matrix for PM2.5 (winter) at LF						
	Predicted					
Actual	d+1 False	d+1 True	d+2 False	d+2 True	d+3 False	d+3 True
False	488	59	476	74	488	63
True	41	144	55	127	65	116

Результаты для зимнего периода

Матрица ошибок для BRF:

Confusion matrix for PM2.5 (winter) at LF

	Predicted					
Actual	d+1 False	d+1 True	d+2 False	d+2 True	d+3 False	d+3 True
False	480	67	461	89	450	101
True	35	**150**	46	**136**	52	**129**

Заключение Если рассматривать только зимний период, характеристики набора данных меняются: с одной стороны, данные более сбалансированы, с другой стороны объем обучающей выборки меньше.

Заключение

- Аппроксимация приращений PM дает лучшие результаты по сравнению с аппроксимацией абсолютных значений.
- Признаки приращения метеорологических переменных повышают качество решения задачи как в постановке классификации, так и в постановке регрессии.
- Признак разности температур улучшает качество прогнозирования до 1%.
- Внедрение шума в данные и перевзвешивание событий выборки повышают качество решения задачи классификации по всем мерам качества.
- Данные за зимний период более сбалансированы, но выборка в два раза меньше, что может приводить к снижению достоверности результатов моделей.

Дальнейшие перспективы: Использование методов глубокого обучения

Заключение

- Аппроксимация приращений PM дает лучшие результаты по сравнению с аппроксимацией абсолютных значений.
- Признаки приращения метеорологических переменных повышают качество решения задачи как в постановке классификации, так и в постановке регрессии.
- Признак разности температур улучшает качество прогнозирования до 1%.
- Внедрение шума в данные и перевзвешивание событий выборки повышают качество решения задачи классификации по всем мерам качества.
- Данные за зимний период более сбалансированы, но выборка в два раза меньше, что может приводить к снижению достоверности результатов моделей.

Дальнейшие перспективы: Использование методов глубокого обучения