Хакатоны как способ апробации open-source фреймворков автоматического машинного обучения

Пинчук Майя

СТУДЕНТ
РАЗРАБОТЧИК AUTOML ФРЕЙМВОРКА FEDOT

Зачем нужно тестировать свои разработки на хакатонах?

- уникальные данные
- активное и плотное использование своего продукта, взгляд со стороны пользователей
- выявление недочетов

Emergency DataHack 2021

ЯСИА

Раньше нормы ожидается вскрытие рек в Якутии

Более 16 тысяч человек привлекут для безопасного пропуска весеннего половодья.

1 день назад

Поставленная задача -- построить предиктивную модель, которая позволит в весенний период оценивать суточное приращение максимального уровня воды на 7 дней вперед на определенных участках реки Лена.

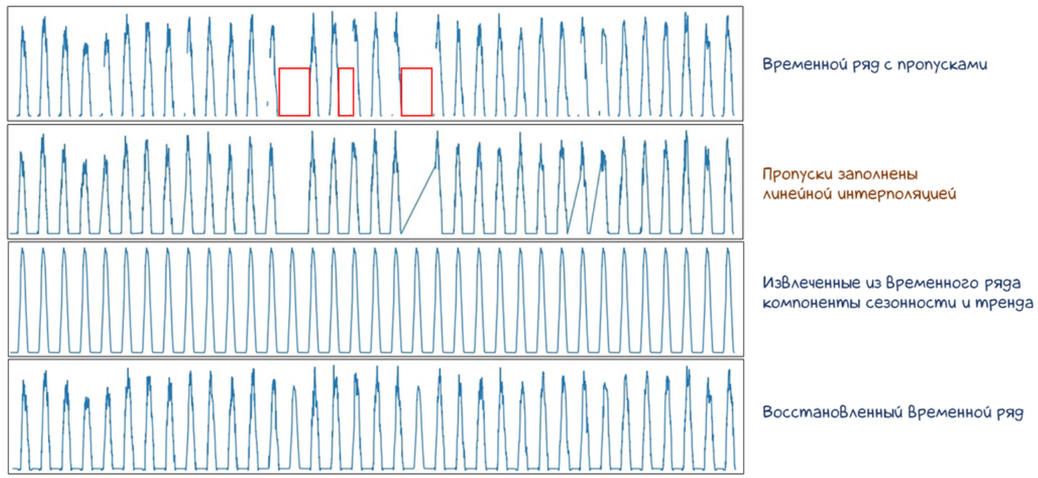
Метрика -- микровзвешенное значение среднеквадратичной ошибки MSE, нормированное на среднеквадратичное отклонение по целевым населенным пунктам.

Формулировка задачи -- регрессия / прогнозирование временного ряда, где целевая переменная - показатель того, насколько вырос уровень воды за сутки.

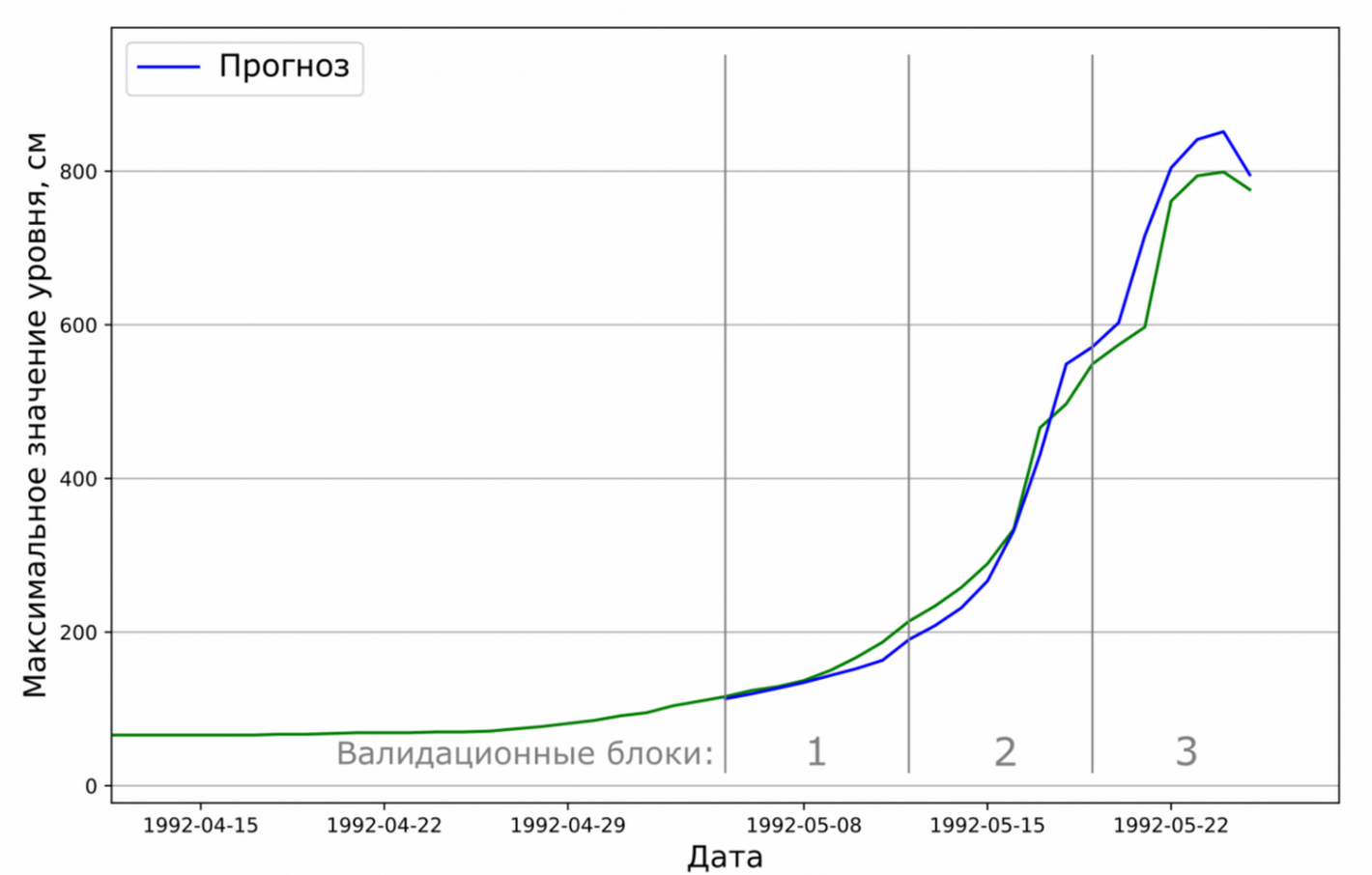
Как был использован FEDOT: заполнение пропусков

Добавление новой функциональности во фреймворк

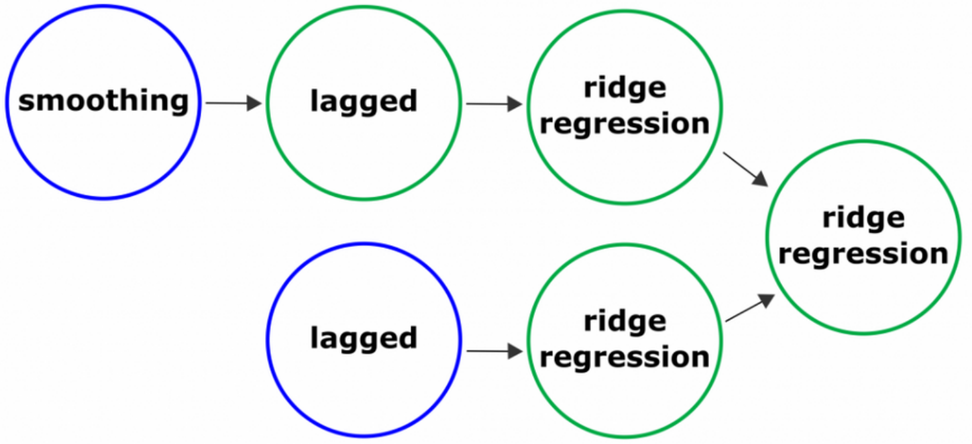
FEDOT



Как был использован FEDOT: предсказание временных рядов



Пример предсказаний уровня во время половодья при помощи модели прогнозирования временного ряда.



Структура используемого при прогнозировании пайплайна

Как был использован FEDOT: многомерная регрессия

Эта модель отвечает за учет метеопараметров и данных о событиях, происходящих на реке.
Планируется решать multi-target регрессию следующего вида:

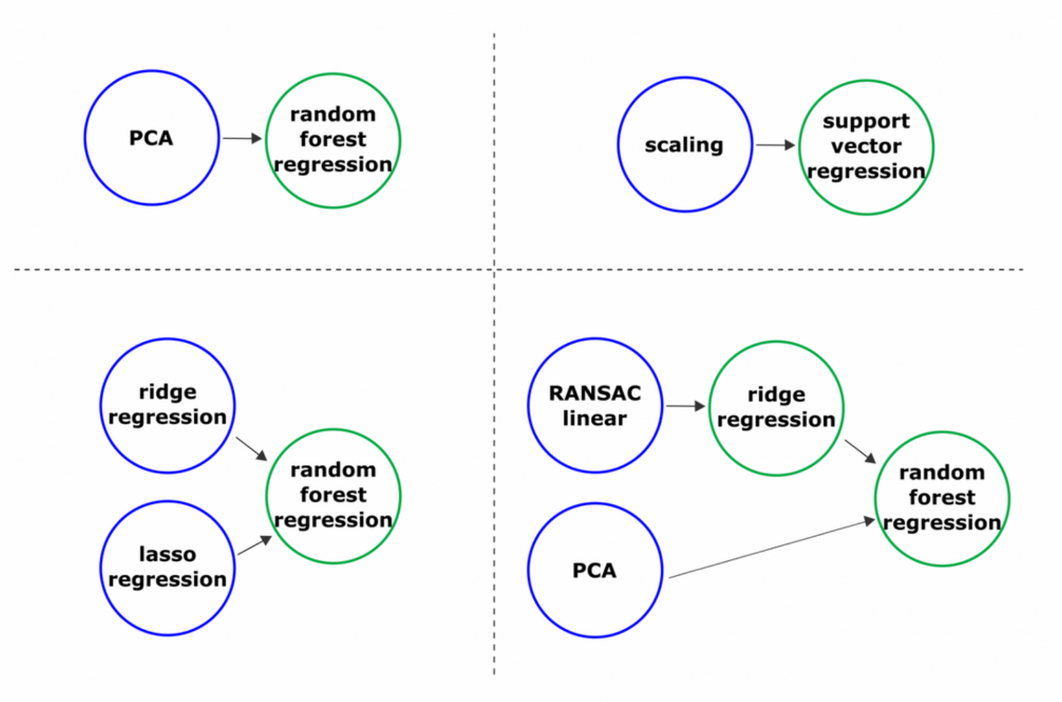
День	Предикторы				Отклики						
					21	22	23	24	25	26	27
20.04.2020	Сумма осадков за 10 суток до данной даты	Амплитуда высоты снежного покрова за 20 суток до данной даты	...	Сумма рангов важности событий за 5 суток							

Целевых переменных в таблице всегда семь, с 21 по 27 апреля в данном примере, что равно горизонту прогнозирования.
Как понятно из описания признаков, они являются производными от исходных метеопараметров.

Как был использован FEDOT: многомерная регрессия

Более обширное тестирование за счет использования разных подходов

После формирования всех признаков на них был запущен FEDOT, чтобы он автоматически определил методы предобработки данных и подобрал модель



Примеры найденных пайплайнов при решении задачи multi-target регрессии

Результаты

Пример прогнозов, получаемых из двух моделей

Место	Команда	Датасет	Скор	Кол-во загрузок
1	NSS_lab team	4	3.2020969195617877	5

Хакатон от Цифрового прорыва

Задача
Построение адаптивного алгоритма для прогнозирования макроэкономических данных

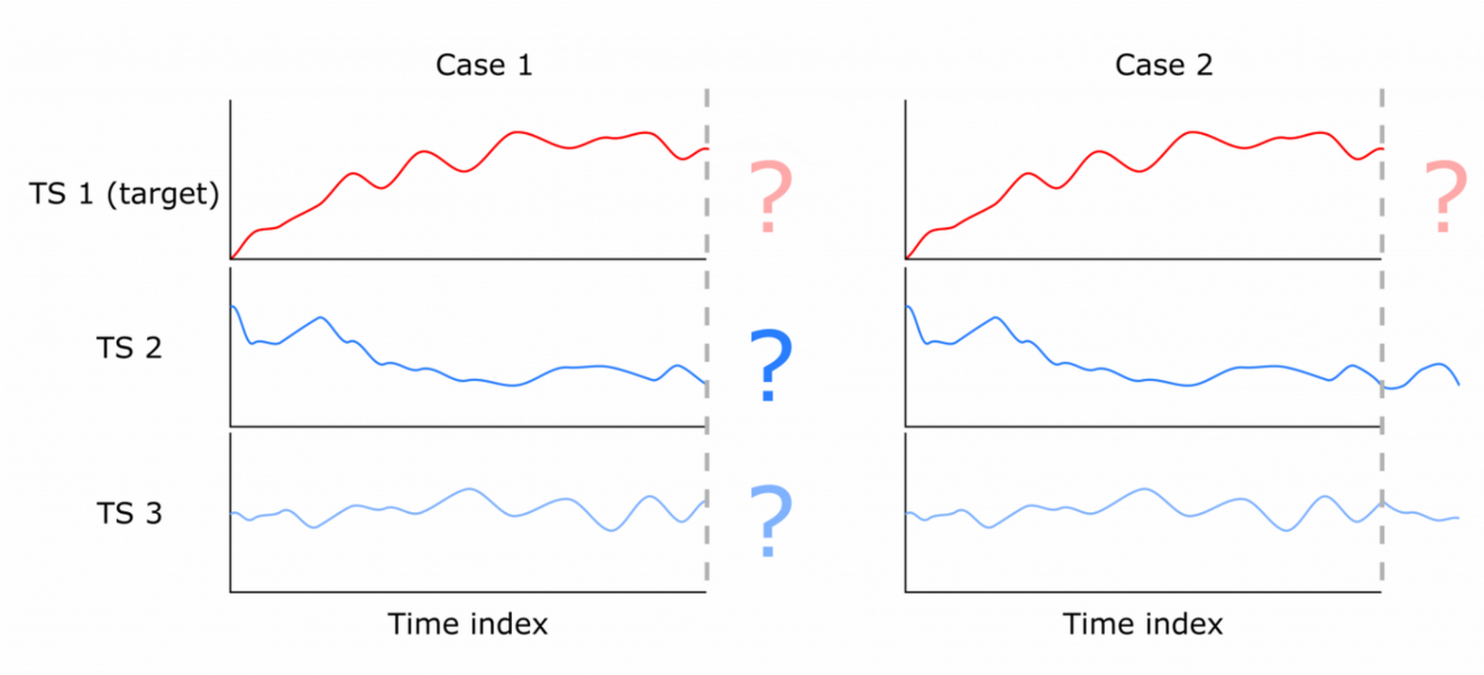
Проблема
Особенностью задач макроэкономического и макрофинансового прогнозирования является работа с наборами данных с относительно небольшим количеством наблюдений, что усложняет процесс построения качественных моделей.

Данные и ход решения

Типы временных рядов:

- Многомерные временные ряды
- Временные ряды с экзогенными переменными

В некоторых временных рядах были пропуски



Многомерные временные ряды и с экзогенной переменной

Как был использован FEDOT: заполнение пропусков

Прогноз на предыстории + прогноз на следующей после пропуска части временного ряда объединяются с помощью взвешенного среднего.

	Диффузный индекс цен на выпускаемую продукцию, ожидаемые изменения	Диффузный индекс цен на покупаемую продукцию, ожидаемые изменения	Диффузный индекс заработной платы, ожидаемые изменения
Преобразование			
2003m01	74,00	93,00	61,00
2003m02	78,00	95,00	66,00
2003m03	84,00	98,00	72,00
2003m04	83,00	95,00	70,00
2003m05	79,00		76,00
2003m06	75,00		73,00
2003m07	74,00		74,00
2003m08	76,00		68,00
2003m09	74,00	90,00	76,00
2003m10	71,00	91,00	70,00
2003m11	77,00	88,00	65,00
2003m12	72,00	91,00	65,00
2004m01	80,00	93,00	64,00
2004m02	81,00	94,00	66,00
2004m03	85,00	94,00	68,00
2004m04	76,00	93,00	70,00

Пример пропуска в данных с хакатона

Как был использован FEDOT: временные ряды с экзогенной переменной

Взгляд со стороны пользователей

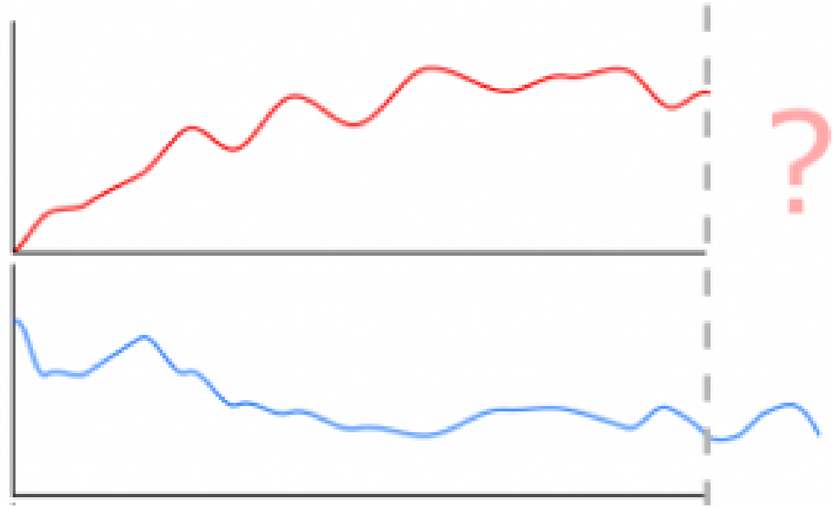
Как предсказание временных рядов с экзогенной переменной работает в FEDOT?

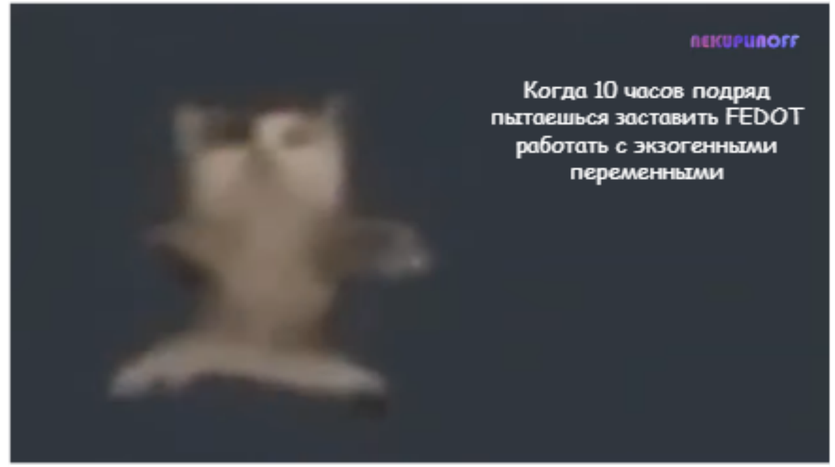
exog_ts	0	1	2	3	4	5
target_ts	1	2	3	4	5	6

	target_ts	target_ts	exog_ts	exog_ts		target_ts	target_ts
features	1	2	2	3	**targets**	3	4

	target_ts	target_ts	exog_ts	exog_ts		target_ts	target_ts
features	2	3	3	4	**targets**	4	5

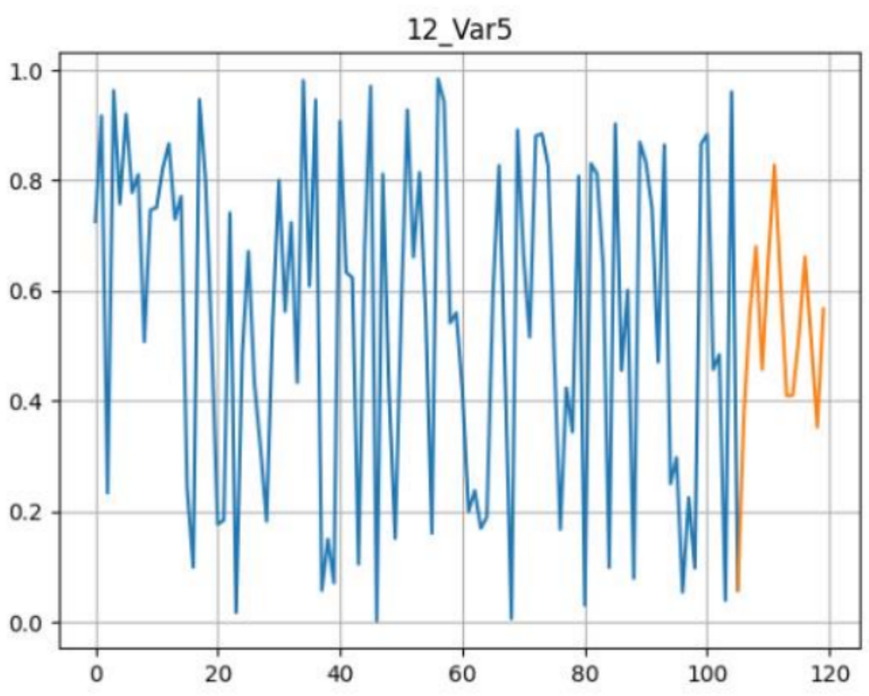
Примеры предсказания с горизонтом предсказния FL=2

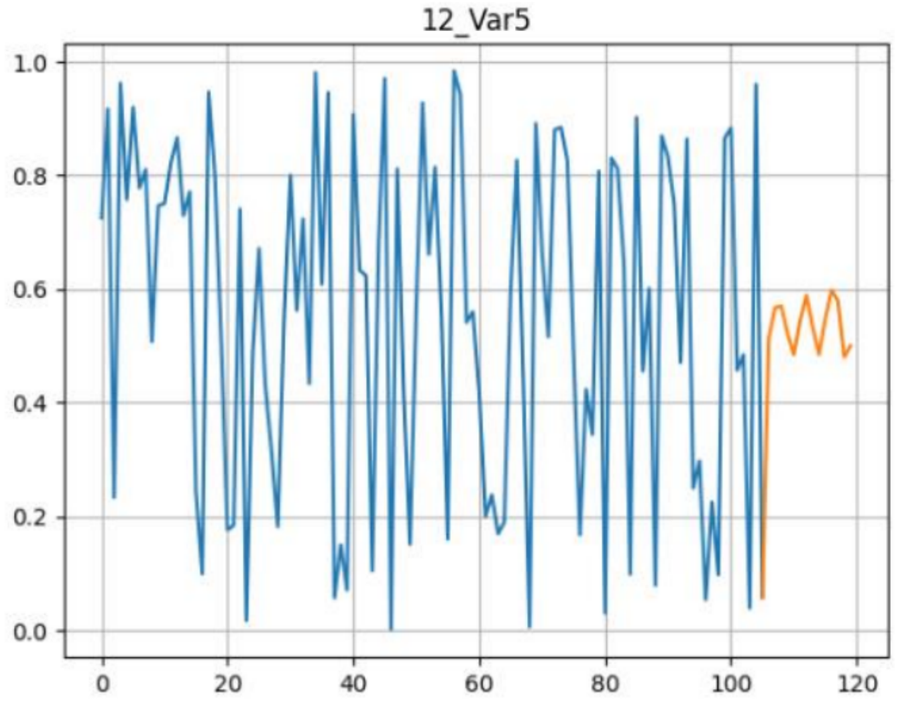




10 ЧАСОВ - ШНИ ШНА ШНАПИ ШНАПИ ШНАПИ

Как был использован FEDOT: временные ряды с экзогенной переменной





Пример предсказания с экзогенной переменной (слева) и без (справа)

Как был использован FEDOT: тюнинг

Что важнее: композирование или тюнинг?

В FEDOT под тюнингом подразумевается настройка гиперпараметров модели

MAPE, %	Var	Std	Mean	Median	Min	Max
Квантили без тюнинга	7856990	28030.3	8476.36	2996.06	98.84	235671
Корреляция без тюнинга	7856990	28030.3	8476.36	2996.06	98.84	235671
Квантили +тюнинг (10 it)	**8.722**	**29.532**	**8.903**	1.894	**0.001**	**240.324**
Корреляция +тюнинг (10 it)	11.612	34.077	11.788	**1.772**	0.009	240.463

Пример получившихся метрик с тюнингом и без

Общая идея и результат

3 место

FEDOT TEAM

Реализация адаптивного алгоритма для прогнозирования данных макроэкономических показателей

Кодировать каждый временной ряд с помощью вектора извлеченных из него признаков (в нашем случае это были квантили) -> использовать для ряда из теста затюненную модель для ближайшего ряда из трейна

MAPE, %	Var	Std	Mean	Median	Min	Max
Без метаобучения Тюнинг (50 it)	0.027	1.648	9.738	9.738	8.090	11.387
Метаобучение+ тюнинг (10 it)	8.722	29.532	8.903	1.894	**0.001**	**240.324**
Метаобучение+ тюнинг (25 it)	**8.713**	**29.518**	**8.374**	**1.313**	0.019	241.450

Код нашего решения доступен здесь:
https://github.com/maypink/digital_breakthrough_hack

Хакатон от «Газпром нефть»

Задача

Поиск алгоритмов и построение самообучающихся математических моделей для прогнозирования составов новых продуктов из нефти.

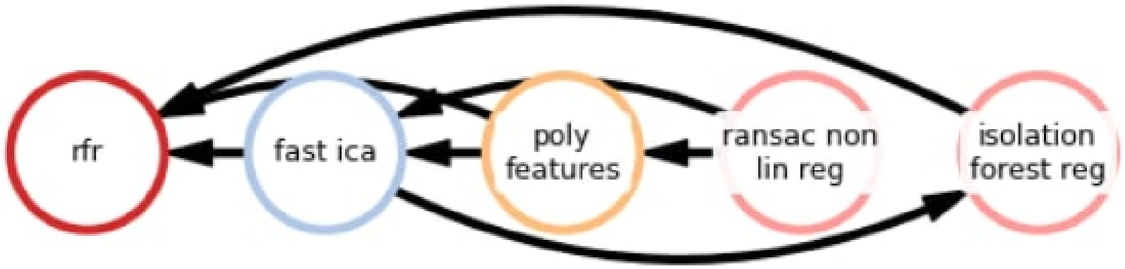
Во время построения модели должны учитываться такие показатели качества, как глубина проникновения иглы, эластичность, растяжимость при различных температурах.

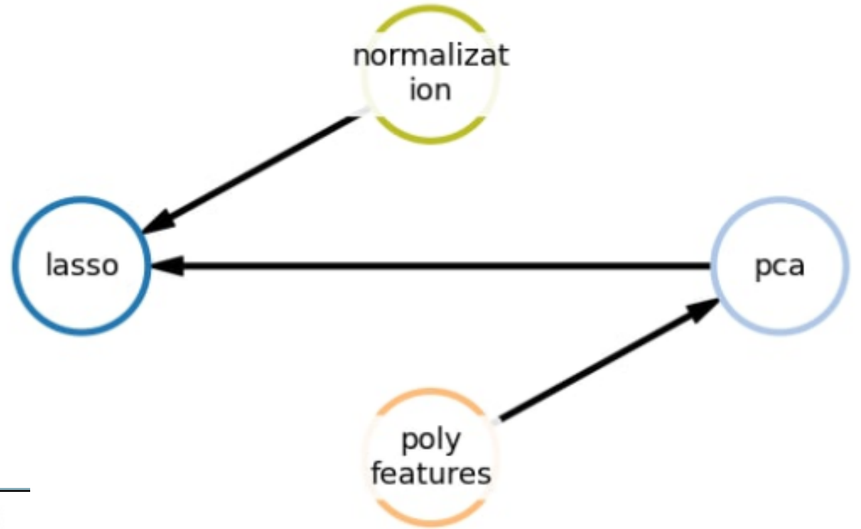
Актуальность

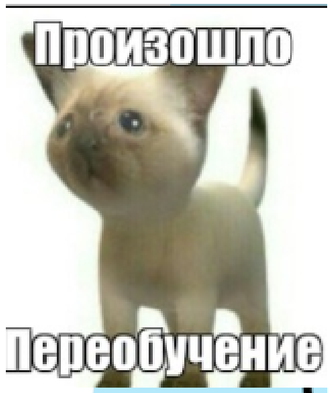
В дальнейшем это сможет привести к более быстрой разработке новых материалов для дорожного покрытия и появлению качественных и долговечных дорог.

Как был использован FEDOT: композирование

Данных было очень мало, поэтому процесс композирования запускать было нецелесообразно







Как был использован FEDOT: инструмент для построения пайплайнов + тюнинг

Использовать FEDOT как ML инструмент

FEDOT пригодился нам для быстрой предобработки данных и удобного построения пайплайнов + тюнинга
Под тюнингом в FEDOT понимается подбор гиперпараметров

```
6       problem = 'classification'
7       train_data_path = f'{fedot_project_root()}/cases/data/scoring/scoring_train.csv'
8       test_data_path = f'{fedot_project_root()}/cases/data/scoring/scoring_test.csv'
9
10      baseline_model = Fedot(problem=problem, timeout=timeout, seed=42)
11      baseline_model.fit(features=train_data_path, target='target', predefined_model='rf')
12
13      baseline_model.predict(features=test_data_path)
14      print(baseline_model.get_metrics())
```

Пример того, что можно запустить FEDOT на своих данных почти забесплатно

Общие идеи решения

Итеративное предсказание может помочь предсказывать более точно из-за имеющейся корреляции между переменными

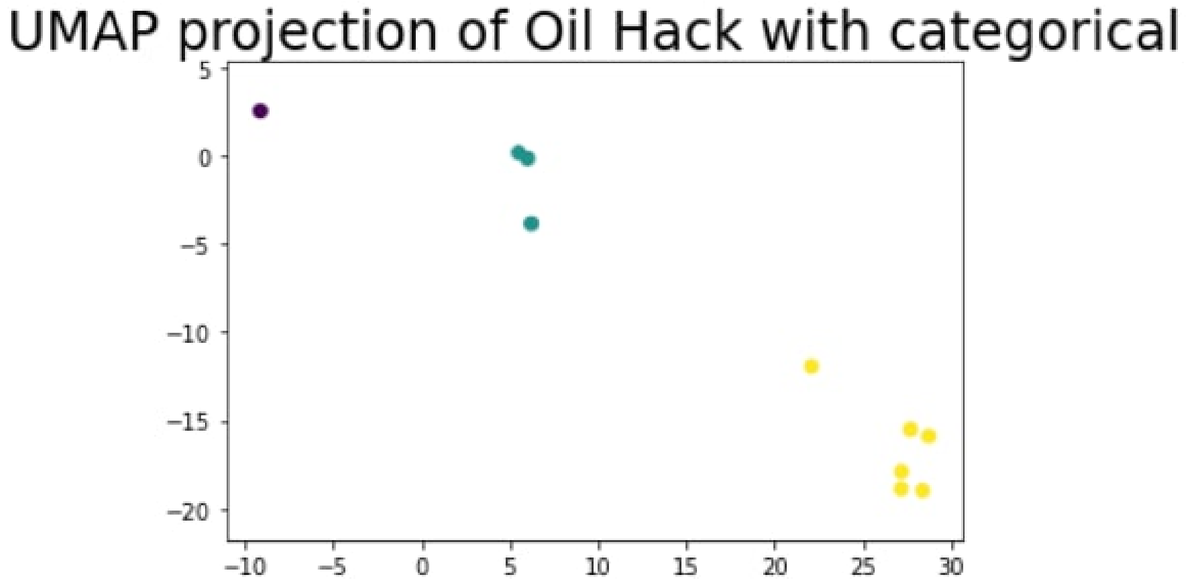
	Глубина проникания иглы при 0 °C, [мм-1]	Глубина проникания иглы при 25 °C, [мм-1]	Растяжимость при температуре 0 °C, [см]	Температура размягчения, [°C]	Эластичность при 0 °C, [%]
Глубина проникания иглы при 0 °C, [мм-1]	1.000000	0.866932	0.626220	-0.389027	0.574053
Глубина проникания иглы при 25 °C, [мм-1]	0.866932	1.000000	0.632999	-0.325991	0.567312
Растяжимость при температуре 0 °C, [см]	0.626220	0.632999	1.000000	-0.187600	0.570063
Температура размягчения, [°C]	-0.389027	-0.325991	-0.187600	1.000000	-0.269470
Эластичность при 0 °C, [%]	0.574053	0.567312	0.570063	-0.269470	1.000000

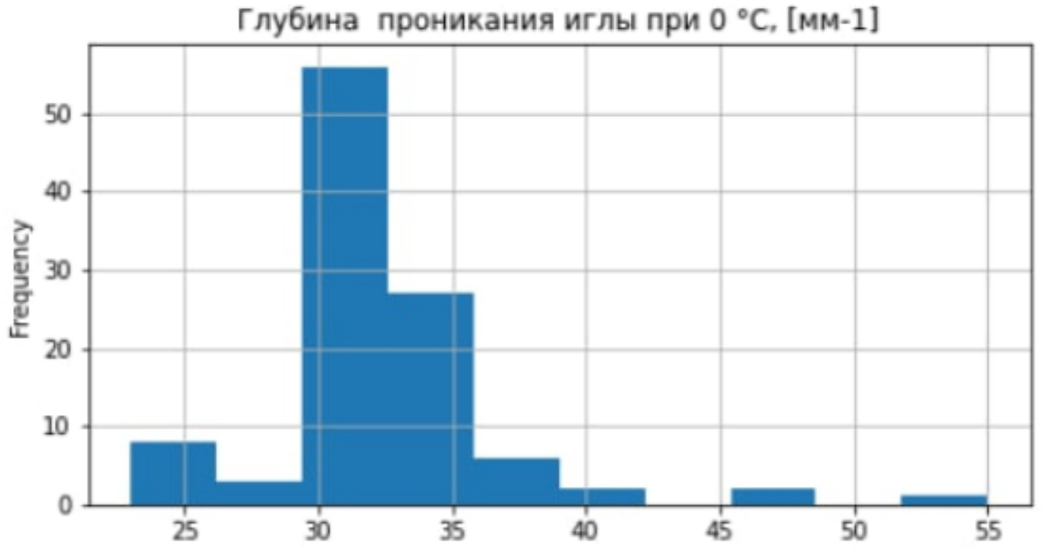
Зависимости между показателями качества и содержанием продукта уже были выявлены, так что можно попробовать закодировать их явно

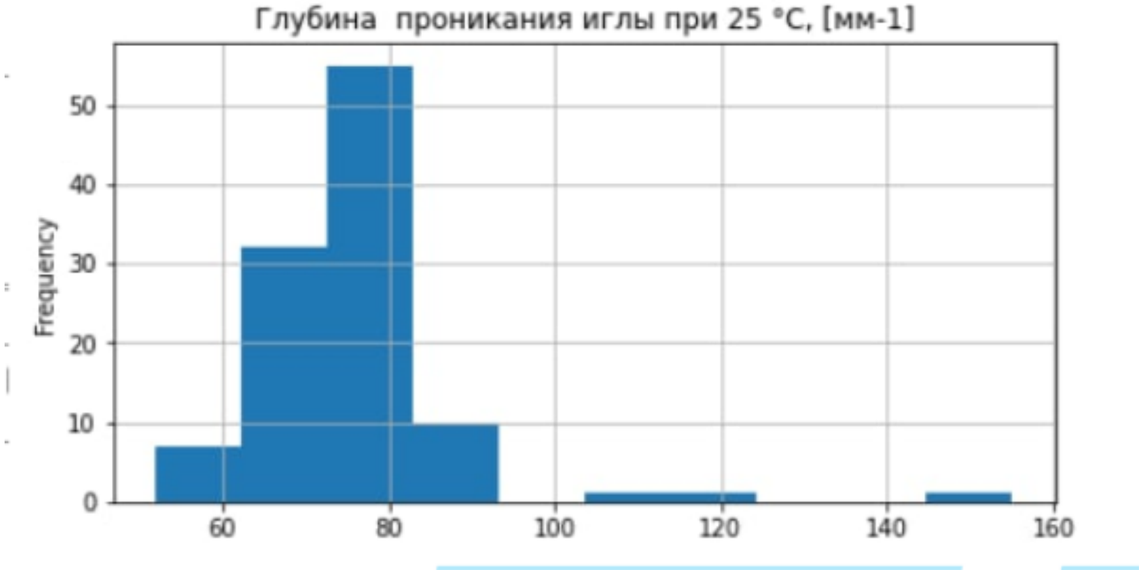
			диапазон			влияние компонентов на показатели качества при увеличении ввода			
Показатели качества	**метод испытания**	**единица измерения**	**минимум**	**максимум**		**пластификатор**	**полимер**	**сшивающая добавка**	**адгезионная добавка**
Глубина проникания иглы при 25 °C	ГОСТ 11501	0,1 мм	60	95		увеличивает	снижает		незначительно снижает

Общие идеи решения

Кластеризация при помощи UMAP помогла выявить три явно различимых кластера.







Общие идеи решения

Кластеризация + тюнинг

В качестве модели был использован следующий пайплайн: полиномиальные признаки+ нормализация + гребневая регрессия.

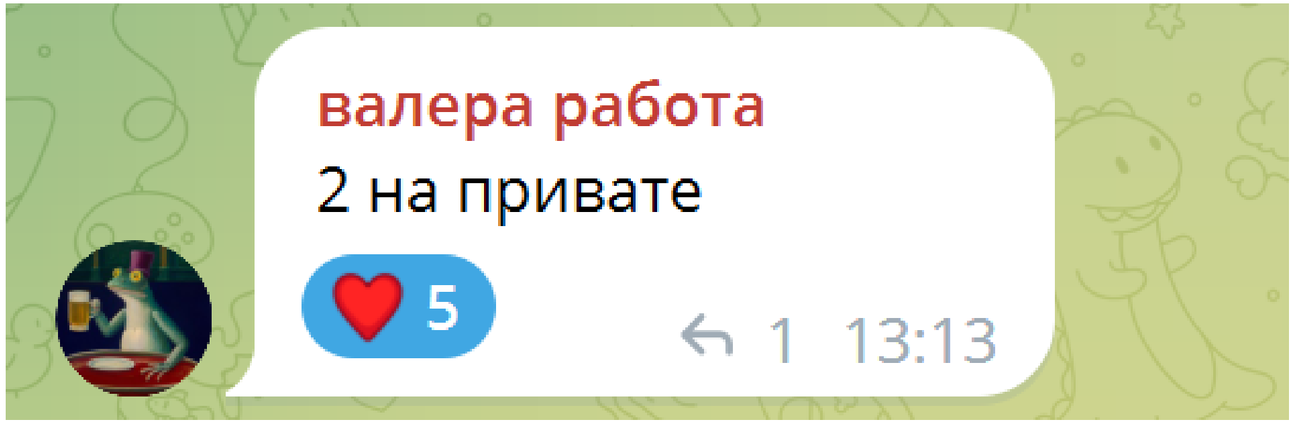
- Простота
- Интерпретируемость
- Эффективность
- Устойчивость

Итоговый пайплайн

Результат

На хакатоне было два лидерборда:
публичный и приватный.
Мы заняли **второе место** и там, и там

Метрика должна минимизироваться...



Хакатон от МФТИ

Задача

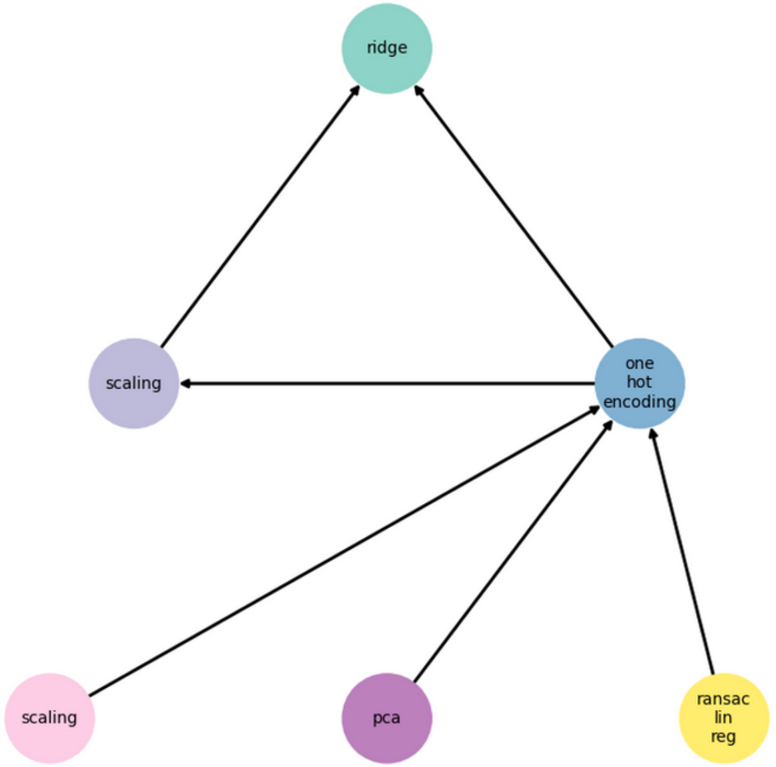
Предсказать потенциальную энергию молекулярной конформации по её атомному составу и координатам атомов

Актуальность

1.Снижение затрат на проведение экспериментов

2.Ускорение синтеза молекул

Как был использован FEDOT: композирование + тюнинг



MAE: 0.04266
Запуск на 4 часа

MAE: 0.06121
Запуск на 8 часов

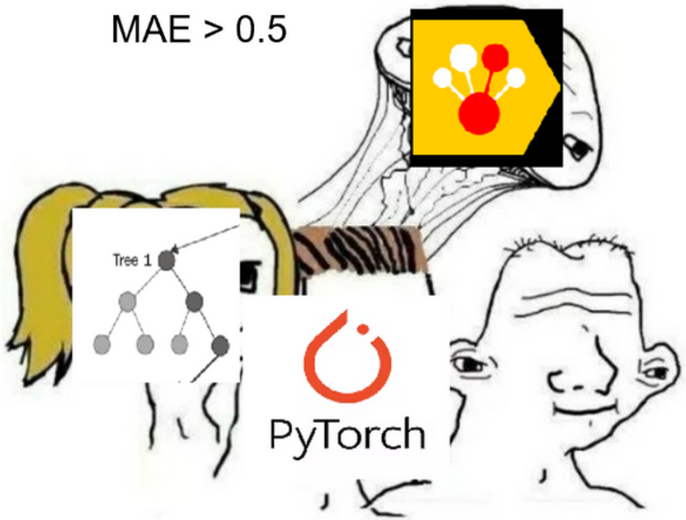
Результат

Было опробовано несколько вариантов:

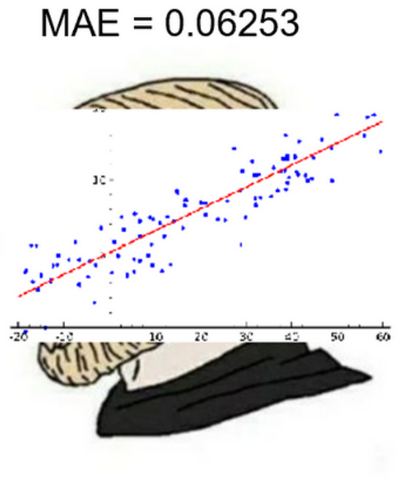
- Линейная регрессия, MAE ~ 0.06
- SchNet, MAE ~ 3.2
- FEDOT, MAE ~ 0.04

Лучший результат с большим отрывом выдал фреймворк FEDOT

В итоге команда заняла 3-е место!

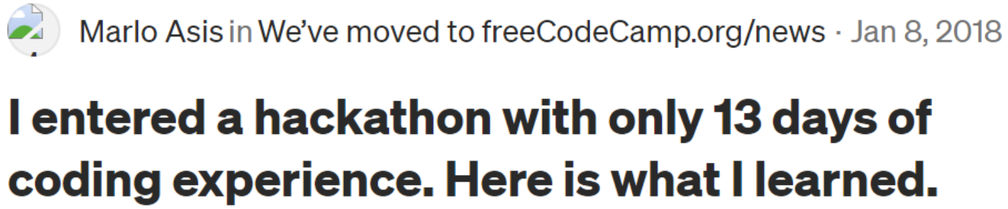


АХАХАХА ТЫ ПРАВДА ЛИНЕЙНЫЙ?



Да

Выводы



- можно использовать AutoML фреймворки (FEDOT) не только для решения непосредственно задачи, но и как удобный инструмент для работы с данными и пайплайнами
- AutoML можно применять (или по крайней мере пробовать) на почти любом хакатоне: так или иначе это даст свои плоды

Спасибо за внимание!