Neoflex

Автогенерация синтетических данных с использованием алгоритмов Generative AI

SmartData 2024

Neoflex

Анастасия Коткова

Ведущий эксперт
команды Data Science R&D

Данила Самошкин

Руководитель направления
внедрения MLOps решений

Какие данные мы будем называть синтетическими?

Данные, которые создаются с помощью искусственных компьютерных алгоритмов с обязательным сохранением их структуры, а также статистических распределений и зависимостей в них

Виды генерируемых данных

Синтетические данные могут имитировать практически любые типы реальных данных

Данные пользователей

Имя, возраст, пол, локация, интересы и предпочтения

Финансовые данные

Счета, история транзакций, балансы, поступления

Медиа контент

Изображения, аудиофайлы, видео

Медицинские данные

История болезни, лечения, хронические диагнозы

Гео-данные

Широта, долгота, высота, гео-метки

Коммерческие данные

Категория товара, SKU, цена, история покупок, отзывы

Обмен данными без ограничений

Реальные датасеты могут включать персональные данные или другую конфиденциальную информацию

Обмен данными без ограничений

Реальные датасеты могут включать персональные данные или другую конфиденциальную информацию

Синтетические данные можно

- Передавать внешним контрагентам
- Делиться со смежными подразделениями внутри компании

Больше данных для обучения и тестирования

Реальных данных может быть недостаточно для качественного обучения и валидации моделей или тестирования приложений

Больше данных для обучения и тестирования

Реальных данных может быть недостаточно для качественного обучения и валидации моделей или тестирования приложений

Синтетические данные позволяют

- Быстро создавать большое количество данных высокого качества
- Эмулировать среды для обучения и тестирования моделей в том случае, когда это невозможно в реальном мире

Сферы применения

Ускоряют внедрение инноваций, упрощают тестирование, обеспечивают рост производительности

Data Science

- Обучение моделей
- Валидация моделей
- Тестирование

R&D

- Разработка концепций
- Тестирование концепций
- Проведение исследований

Development & QA

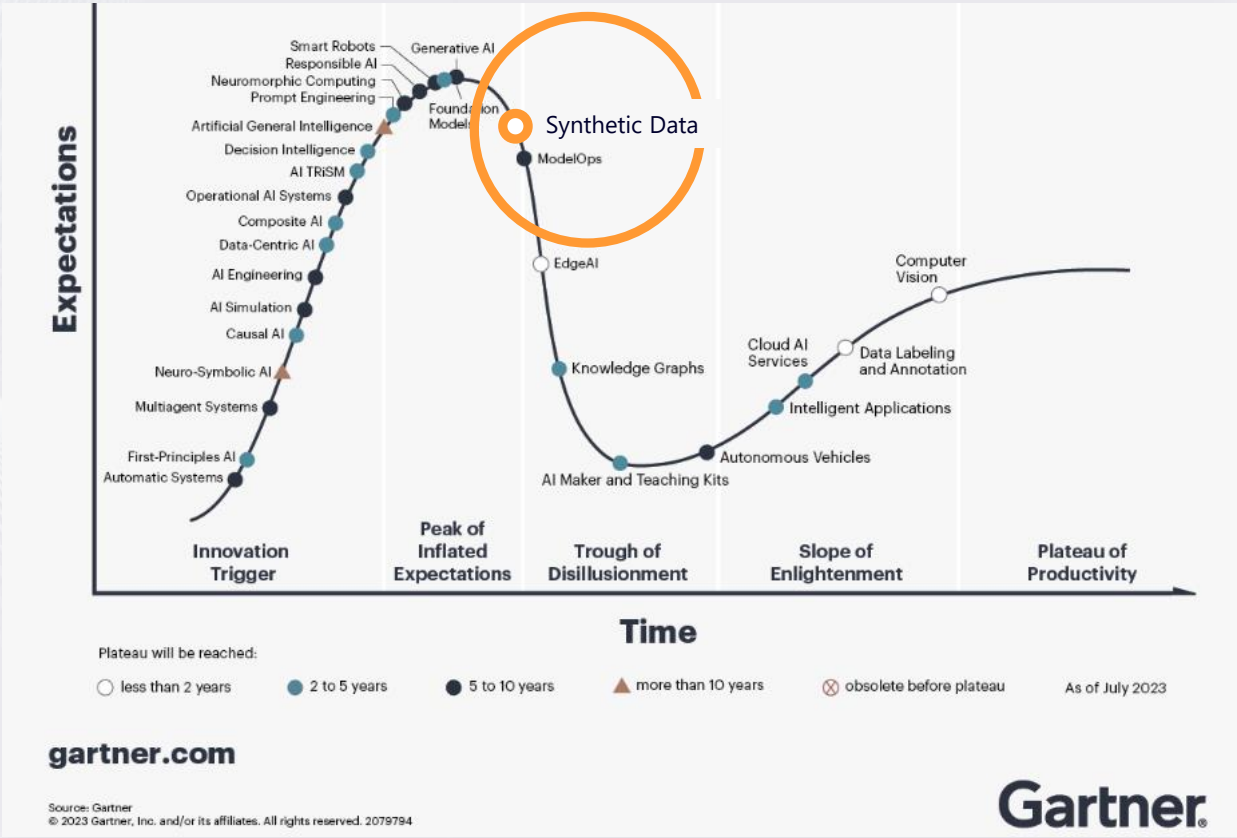
- Функциональное тестирование
- Нагрузочное тестирование

BI

- Доработка наглядности отчетов
- Анализ трендов в данных

Синтетические данные на пике ожиданий

Использование синтетических данных активно внедряется в процессы компаний



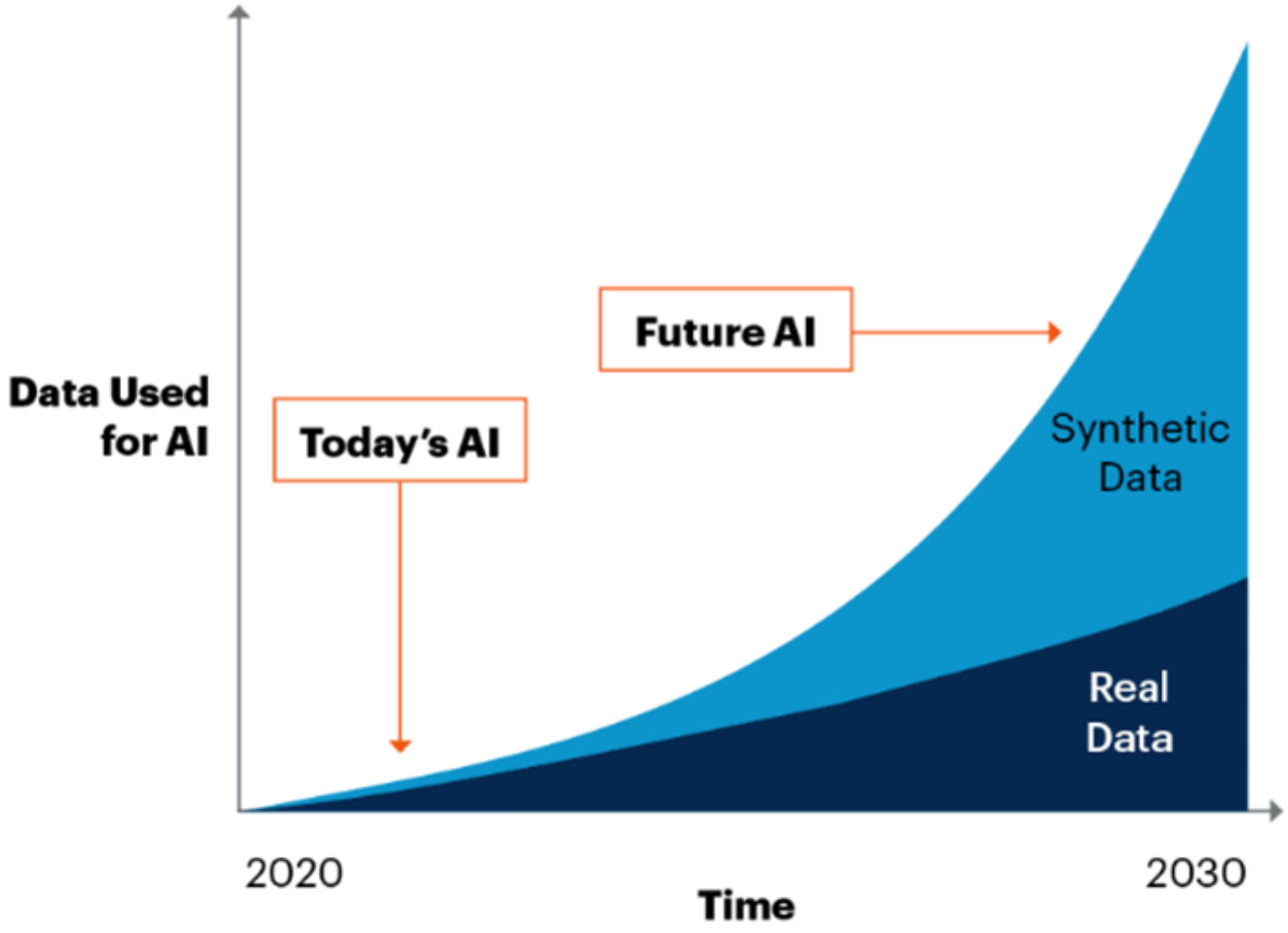
Крупнейшие аналитические агентства отмечают высокий интерес к использованию синтетических данных

[Gartner: Is Synthetic Data the Future of AI? (gartner.com)]

Актуальность подтверждается множеством отчетов

Gartner:

«Благодаря последним достижениям в области AI, синтетические данные теперь могут соответствовать реальным данным по качеству или даже превосходить их»



Методы генерации

Существует несколько отличающихся подходов по генерации данных

Инструменты генерации синтетических данных

Подходы к статической генерации

- Генерация данных на основе правил
- Параметрические модели
- Линейная интерполяция
- Случайная выборка

Модели на основе машинного обучения

- GAN (CTGAN)
- VAE (TVAE)
- Gaussian Copula
- LSTM
- Автоэнкодеры
- RNN

Hybrid методы

Это методы, которые комбинируют правила и модели для генерации данных

А как же оценить качество?

Точность

- Статистическое сходство
- Тест Колмогорова-Смирнова
- Полнота категории и диапазона
- Сохранение границ
- Неполное сходство данных
- Коэффициент корреляции и случайности

Полезность

- Оценка прогнозирования
- Feature importance score
- QScore

Конфиденциальность

- Оценка точного совпадения
- Новизна строк
- Коэффициент вероятности правильной атрибуции
- Вывод
- Выделение
- Связываемость

Преимущества синтетических данных

	Синтетические данные	Реальные данные
Объем	Данные практически любого необходимого объема	Количества имеющихся наборов данных может быть недостаточно
Доступность	Доступ без ограничений, быстрая и простая генерация	Сбор данных может быть долгим или вообще невозможным
Качество	Данные высокого качества в соответствии с требованиями к ним	Качество данных может быть весьма низким
Защита	Нет риска раскрытия персональных или чувствительных данных	Содержат персональную и конфиденциальную информацию

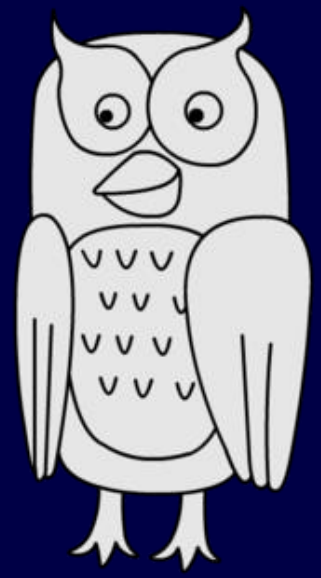
При чем здесь мы?

Предпосылки создания генератора

Требовались данные для тестирования приложений, стримлетов и миграций БД	• GPT модели для генерации таблиц с заданными типами столбцов • Библиотека SDV, если требуется генерация связанных таблиц
Требовался инструмент для безопасной передачи данных между подразделениями	• Использование Faker, Mimesis для анонимизации реальных данных • Создание генератора данных на основе логических функций с заданными распределением величин
Требовались данные для улучшения обучения ML моделей	• Приемы семплирования

ЭТО НЕ РЕШЕНИЕ
А нужно решение!

Давайте внедрять Generative AI!



Наши наблюдения

Результаты ряда интересных примеров
и внутренних экспериментов

Synthetic Data Vault

Python-библиотека для создания синтетических данных, использующая алгоритмы машинного обучения. Сгенерированные данные сохраняют аналогичные характеристики и статистику, как и оригинал

SDV
The Synthetic Data Vault

Подготовьте собственную модель генерации

Набор встроенных ML-моделей для генерации, как одной таблицы, так и целого набора связанных

Оцените качество синтетических данных

Наличие набора встроенных механизмов оценки синтетических датасетов (real vs synthetic)

Gaussian Copula Synthesizer

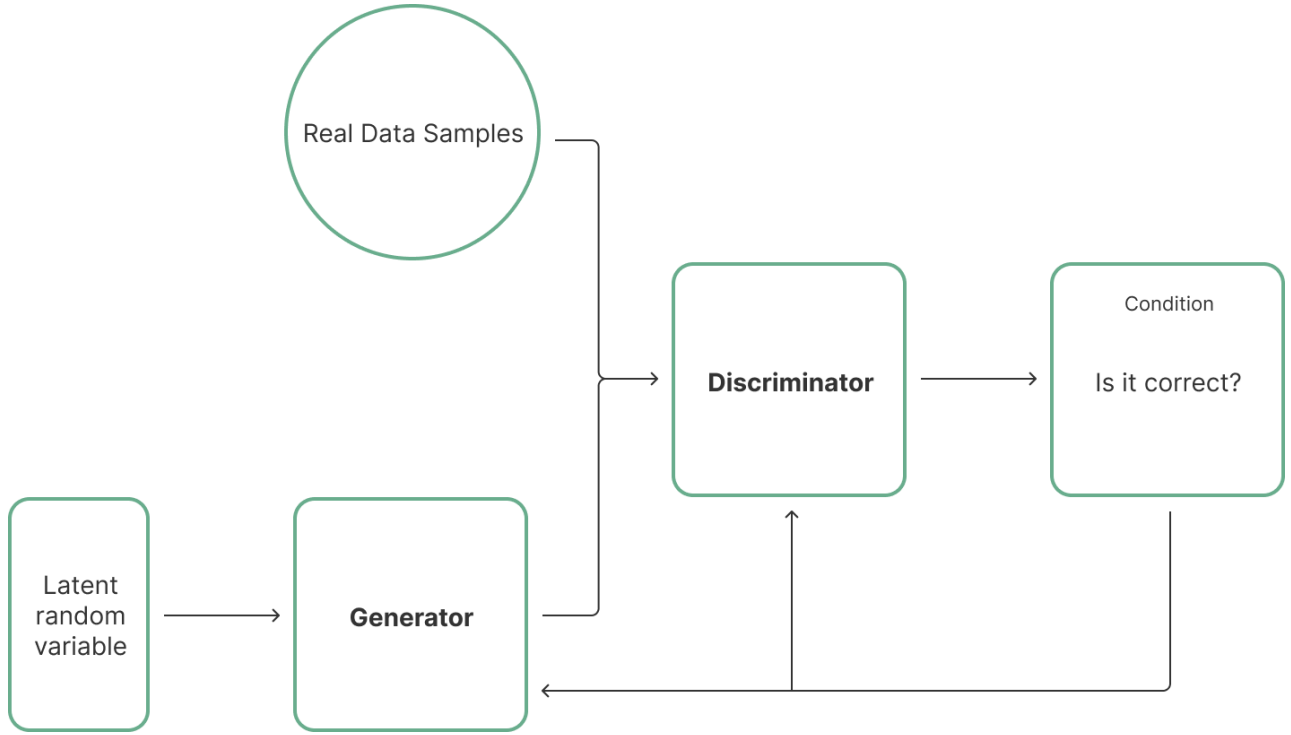
Метод использует статистическую методологию для генерации реалистичных синтетических данных с желаемыми свойствами, такими как нормальное распределение

Используется для данных, которые имеют дискретное распределение, например, вероятность возникновения определенных событий

CTGAN Synthesizer

Генеративно-состязательные сети (generative adversarial network), или GAN — это класс самых популярных моделей для синтезирования данных

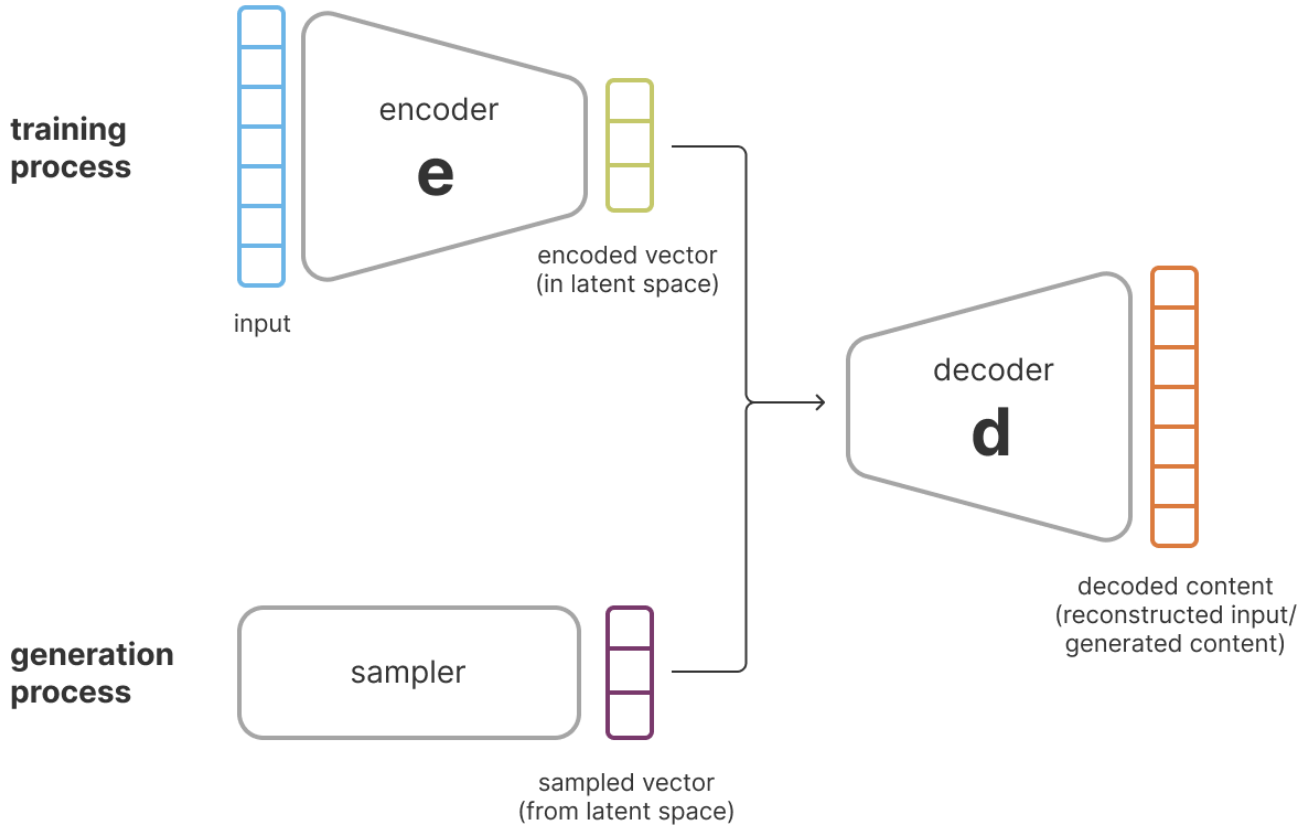
Они состоят из двух подмоделей: генератора и дискриминатора. Задача генератора заключается в синтезе фальшивых данных, а цель дискриминатора — определить, выглядят ли они фальшивыми или реальными. Две модели работают друг против друга, отсюда и состязательность в названии



TVAE Synthesizer

Вариационные автоэнкодеры (variational autoencoder) (VAE) специализируются на определении зависимостей в датасете. Они воссоздают примеры данных из датасета, но в то же время генерируют новые вариации

Вариационные автоэнкодеры используются для генерации различных типов сложных данных, например, рукописного текста, лиц, изображений и табличных данных



Описание данных

	Количество строк			
Data type	**Covid**	**Salaries**	**Credit**	R – Реальные данные S – Синтетические данные
Real data	250 000	2 639	50 000	R
Small synt data	125 000	1 320	25 000	½*S
Mix small data	375 000	3 959	75 000	½*S+R
Synt data	250 000	2 639	50 000	S
Mix data	500 000	5 278	100 000	R+S
Large synt data	500 000	5 278	100 000	2S
Large mix data	750 000	7 917	150 000	2S+R

Анализ моделей генерации для задачи классификации

Лучший метод генерации

Data type	Метрика качества	Covid	Salaries	Credit
Synt data	Precision	GAN	GAN	GAN
	Recall	GAN	GAN	GAN
	F1score	GAN	GAN	COPULA/GAN
Small synt data	Precision	COPULA	COPULA/GAN	TVAE
	Recall	GAN	COPULA/GAN	TVAE
	F1score	COPULA/GAN	COPULA/GAN	TVAE
Mix small data	Precision	GAN	TVAE	GAN
	Recall	GAN	TVAE	EQUALS
	F1score	GAN	TVAE	COPULA/GAN
Mix data	Precision	COPULA/TVAE	EQUALS	TVAE
	Recall	GAN	TVAE/GAN	TVAE
	F1score	COPULA/GAN	TVAE/GAN	TVAE
Large synt data	Precision	TVAE	GAN	GAN
	Recall	GAN	GAN	EQUALS
	F1score	GAN	GAN	COPULA/GAN
Large mixed data	Precision	COPULA	COPULA/GAN	COPULA/GAN
	Recall	GAN	COPULA/GAN	TVAE
	F1score	EQUALS	COPULA/GAN	TVAE

Качество при использовании синтетики в процессе обучения для классификации

Data type	Метрика качества	Covid	Salaries	Credit
Real data	Precision	79	100	26
	Recall	67	100	23
	F1score	71	100	23
Synt data	Precision	78 (-1.3%)	79 (-21%)	27 (+3.9%)
	Recall	67 (0%)	67 (-33%)	21 (-8.7%)
	F1score	71 (0%)	66 (-34%)	18 (-5%)
Mix small data	Precision	82 (+3.8%)	100 (0%)	96 (+269.3%)
	Recall	76 (+13.4%)	100 (0%)	65 (+182.6%)
	F1score	78 (+7%)	100 (0%)	75 (+52.0%)
Mix data	Precision	82 (+3.8%)	100 (0%)	96 (+269.3%)
	Recall	75 (+11.9%)	100 (0%)	69 (+200.0%))
	F1score	77 (+6%)	100 (0%)	79 (+56.0%)
Large mixed data	Precision	84 (+6.3%)	100 (0%)	94 (+261.5%)
	Recall	73 (+9.0%)	100 (0%)	71 (+208.7%)
	F1score	76 (+5%)	100 (0%)	79 (+56.0%)

Анализ моделей генерации для задачи регрессии

Лучший метод генерации

Data type	Метрика качества	Covid	Salaries	Credit
Synt data	Mean_sq_err	TVAE/GAN	TVAE	COPULA
	Mean_abs_err	TVAE/GAN	TVAE	COPULA
	Fr2core	GAN	TVAE	COPULA
Small synt data	Mean_sq_err	COPULA	TVAE	COPULA
	Mean_abs_err	EQUALS	TVAE	TVAE
	Fr2core	COPULA	TVAE	COPULA
Mix small data	Mean_sq_err	GAN	TVAE	COPULA
	Mean_abs_err	TVAE/GAN	TVAE	COPULA
	Fr2core	GAN	TVAE	COPULA
Mix data	Mean_sq_err	COPULA	TVAE	COPULA
	Mean_abs_err	TVAE/GAN	TVAE	COPULA
	Fr2core	COPULA	TVAE	COPULA/TVAE
Large synt data	Mean_sq_err	TVAE/GAN	TVAE	COPULA
	Mean_abs_err	TVAE/GAN	TVAE	COPULA
	Fr2core	GAN	TVAE	COPULA
Large mixed data	Mean_sq_err	EQUALS	TVAE	TVAE
	Mean_abs_err	TVAE/GAN	TVAE	TVAE
	Fr2core	TVAE	TVAE	TVAE

Качество при использовании синтетики в процессе обучения для регрессии

Data type	Метрика качества	Covid	Salaries	Credit
Real data	Mean_sq_err	0.06	39 881 093	1 872 463 747
	Mean_abs_err	0.12	3 974	25 602
	Fr2core	36.82%	96.97%	98.82 %
Synt data	Mean_sq_err	0.06 (0%)	262 957 388 (-559%)	4 071 299 877 (-117%)
	Mean_abs_err	0.13 (-16.7%)	6 641 (-67%)	47 604 (-86%)
	Fr2core	34.21 (-2.6%)	80.02 (-17%)	97.43 (-1.4%)
Mix small data	Mean_sq_err	0.05 (+16.6%)	29 118 470 (+27.0%)	1 023 659 190 (+45.3%)
	Mean_abs_err	0.11 (+8.3%)	2 823 (+28.9%)	17 685 (+30.9%)
	Fr2core	44.8 (+8.0%)	97.79 (+0.8%)	99.35 (+0.5%)
Mix data	Mean_sq_err	0.05 (+16.6%)	23 132 970 (+42.0%)	972 018 697 (+48.1%)
	Mean_abs_err	0.11 (+8.3%)	2 612 (+34.3%)	17 297 (+32.4%)
	Fr2core	44.65 (+7.8%)	98.24 (+1.3%)	99.39 (+0.5%)
Large mixed data	Mean_sq_err	0.06 (+0%)	28 974 132 (+27.3%)	961 451 434 (+48.7%)
	Mean_abs_err	0.11 (+8.3%)	2 772 (+30.3%)	16 950 (+33.8%)
	Fr2core	44.18 (+7.4%)	97.8 (+0.8%)	99.39 (+0.5%)

Что же мы решили сделать?

AutoML процесс

Комплексный сценарий на основе платформы «Neoflex Dognauts» с целью генерации данных при помощи генеративных моделей

NEOFLEX DOGNAUTS

Инструменты генерации синтетических данных

Подходы к статической генерации

- Генерация данных на основе правил
- Параметрические модели
- Линейная интерполяция
- Случайная выборка

Модели на основе машинного обучения

- GAN (CTGAN)
- VAE (TVAE)
- Gaussian Copula
- LSTM
- Автоэнкодеры
- RNN

Hybrid методы

Это методы, которые комбинируют правила и модели для генерации данных

MLFlow, Seldon и Airflow

MLFlow

Для хранения и передачи моделей генерации в реестре моделей

Seldon

Для деплоя и онлайн генерации данных

Airflow

Для запуска пайплайна обучения генератора и пакетной генерации данных

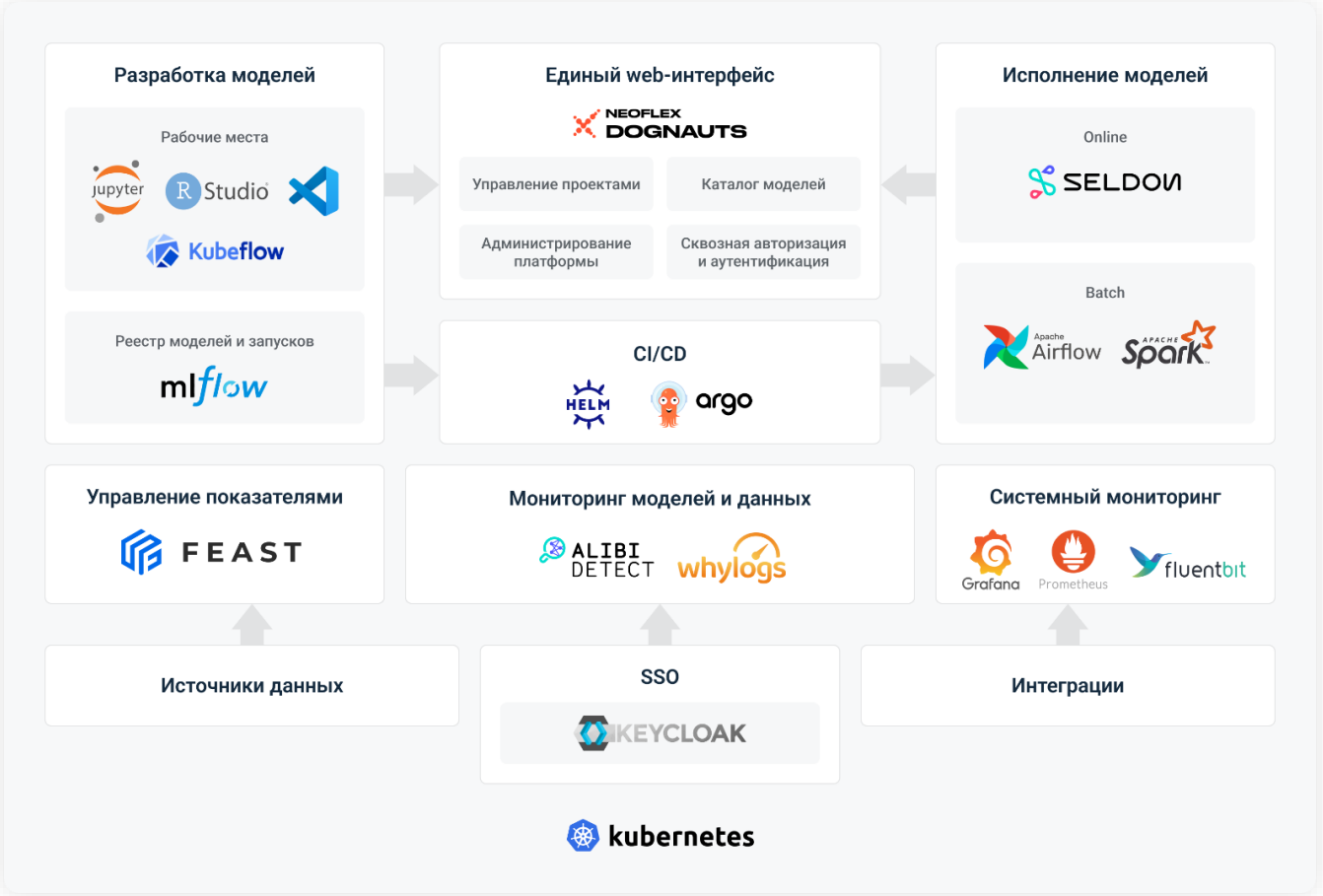
NEOFLEX DOGNAUTS

MLOps-платформа, обеспечивающая полный цикл разработки и эксплуатации моделей машинного обучения

Платформа зарегистрирована в Реестре Российского ПО

Платформа с открытым программным кодом

Архитектура платформы



А как же оценить качество?

Точность

- Статистическое сходство
- Тест Колмогорова-Смирнова
- Полнота категории и диапазона
- Сохранение границ
- Неполное сходство данных
- Коэффициент корреляции и случайности

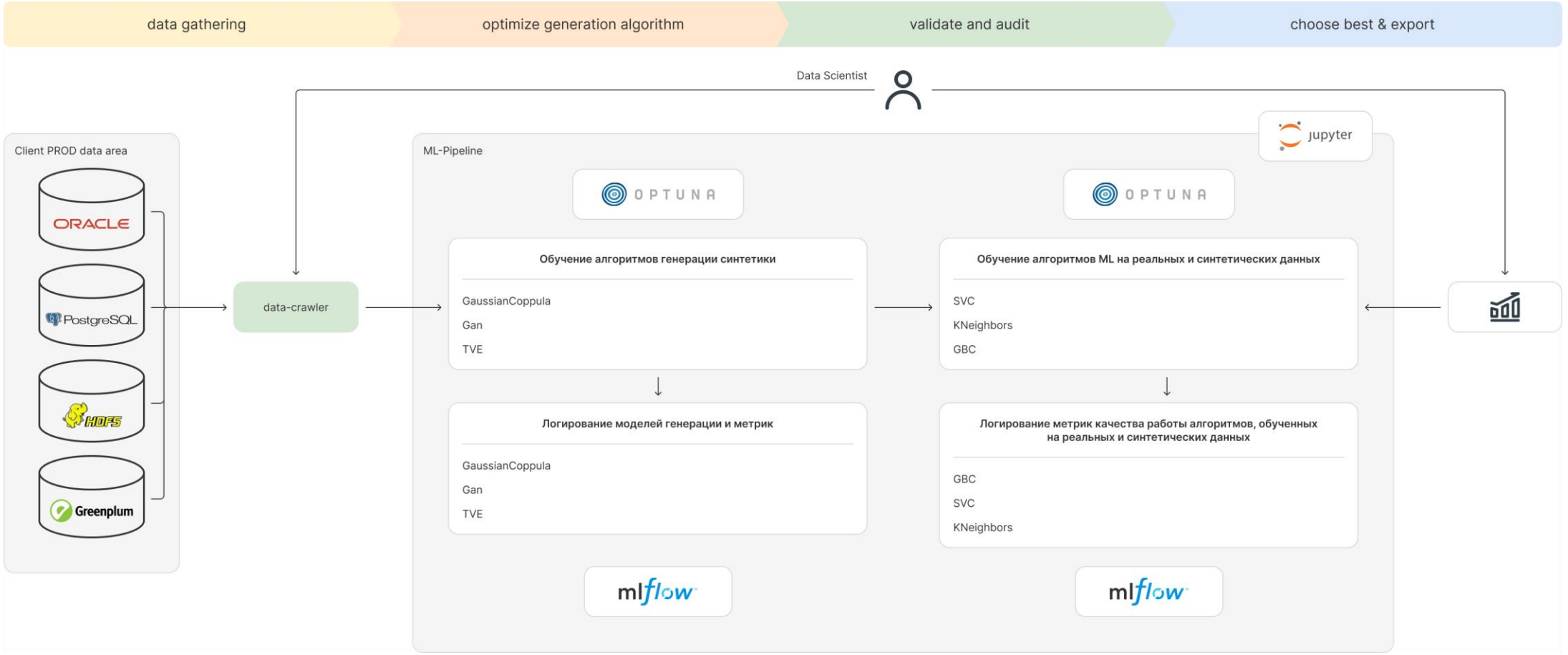
Полезность

- Оценка прогнозирования
- Feature importance score
- QScore

Конфиденциальность

- Оценка точного совпадения
- Новизна строк
- Коэффициент вероятности правильной атрибуции
- Вывод
- Выделение
- Связываемость

Архитектура решения



То, что получилось в итоге

Демо процесса

OPTUNA

Обучение алгоритмов генерации синтетики

GaussianCoppula

Gan

TVE

↓

Логирование моделей генерации и метрик

GaussianCoppula

Gan

TVE

mlflow

```
config = {
    "synthetic_algorithms": ['GaussianCopulaSynthesizer', 'CTGANSynthesizer', 'TVAESynthesizer'],
    "gc_default_distribution": ["norm", "beta", "truncnorm", "uniform", "gamma"],
    "ctg_optimization_sampler": "GridSampler",
    "syn_percent_min": 0.5,
    "syn_percent_max": 1.1,
}

neogen(data_config)
neogen_classification(data_config, gen_config=config)
neogen_regression(data_config, gen_config=config)
```

Synthetic data algorithms optimization

Provide Feedback | Add Description

metrics.rmse < 1 and params.model = "tree" | Time created | State: Active

Group by

Table | Chart | Evaluation Experimental | Traces Experimental

Run Name	Created	Dataset
All algorithms	2 days ago	-
TVAESynthesizer...	2 days ago	-
CTGANSynthesiz...	2 days ago	-
GaussianCopula...	2 days ago	-

Synthetic data algorithms optimization ›

CTGANSynthesizer optimization

Overview | Model metrics | System metrics | Artifacts

Description

No description

Details

Created at	2024-08-23 10:37:54
Created by	GANS
Experiment ID	216741489927507135
Status	Finished
Run ID	9c37446afea943ed9ca9661bf69af2d4
Duration	7.6s
Parent run	All algorithms
Datasets used	—
Tags	Add
Source	D:\Projects\neogen_venv\Lib\site-packages\ipykernel_launcher.py
Logged models	pyfunc
Registered models	—

Parameters (3)

Search parameters

Parameter	Value
enforce_min_max_values	True
enforce_rounding	True
epochs	100

Metrics (1)

Search metrics

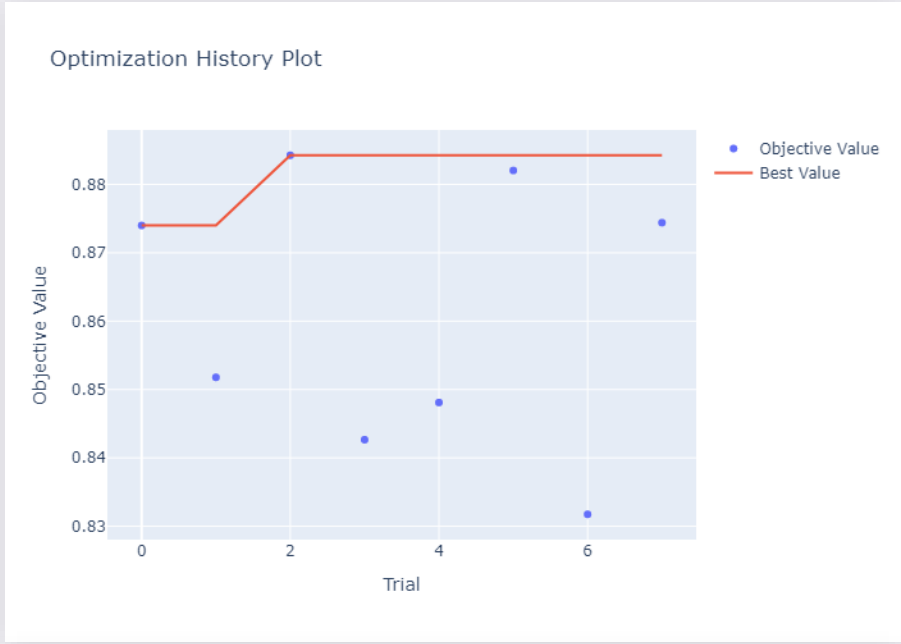
Metric	Value
mean_quality	0.8842594710596856

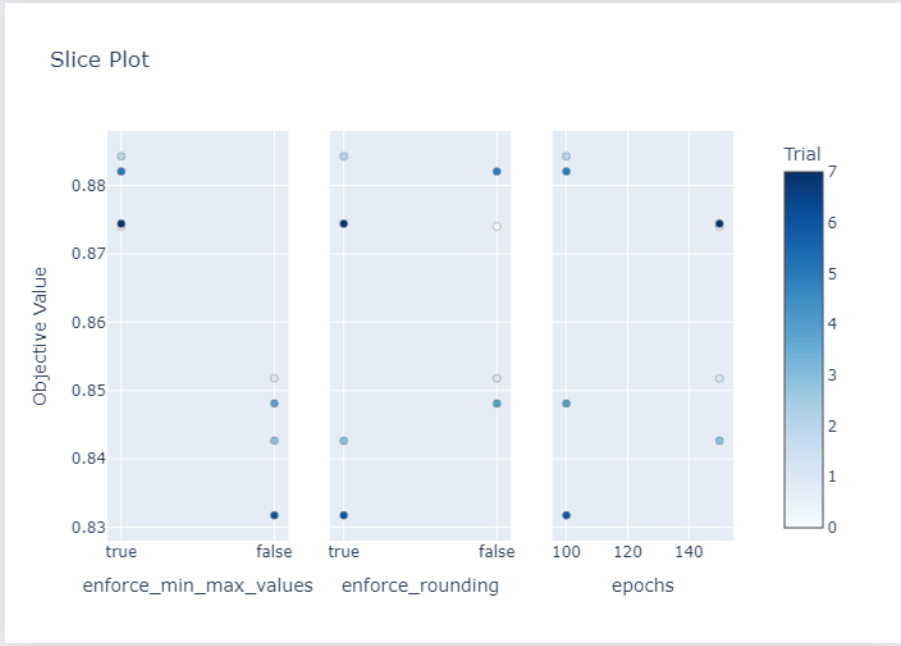
Synthetic data algorithms optimization ›

CTGANSynthesizer optimization

Overview | Model metrics | System metrics | Artifacts

- model
  - artifacts
    - CTGANSynthesizer_synthesizer.pkl
    - encoder.bin
  - MLmodel
  - conda.yaml
  - python_env.yaml
  - python_model.pkl
  - requirements.txt
- optimization
  - all_trials.csv
  - optimization_history.png
  - slice.png





OPTUNA

Обучение алгоритмов ML на реальных и синтетических данных

SVC

KNeighbors

GBC

↓

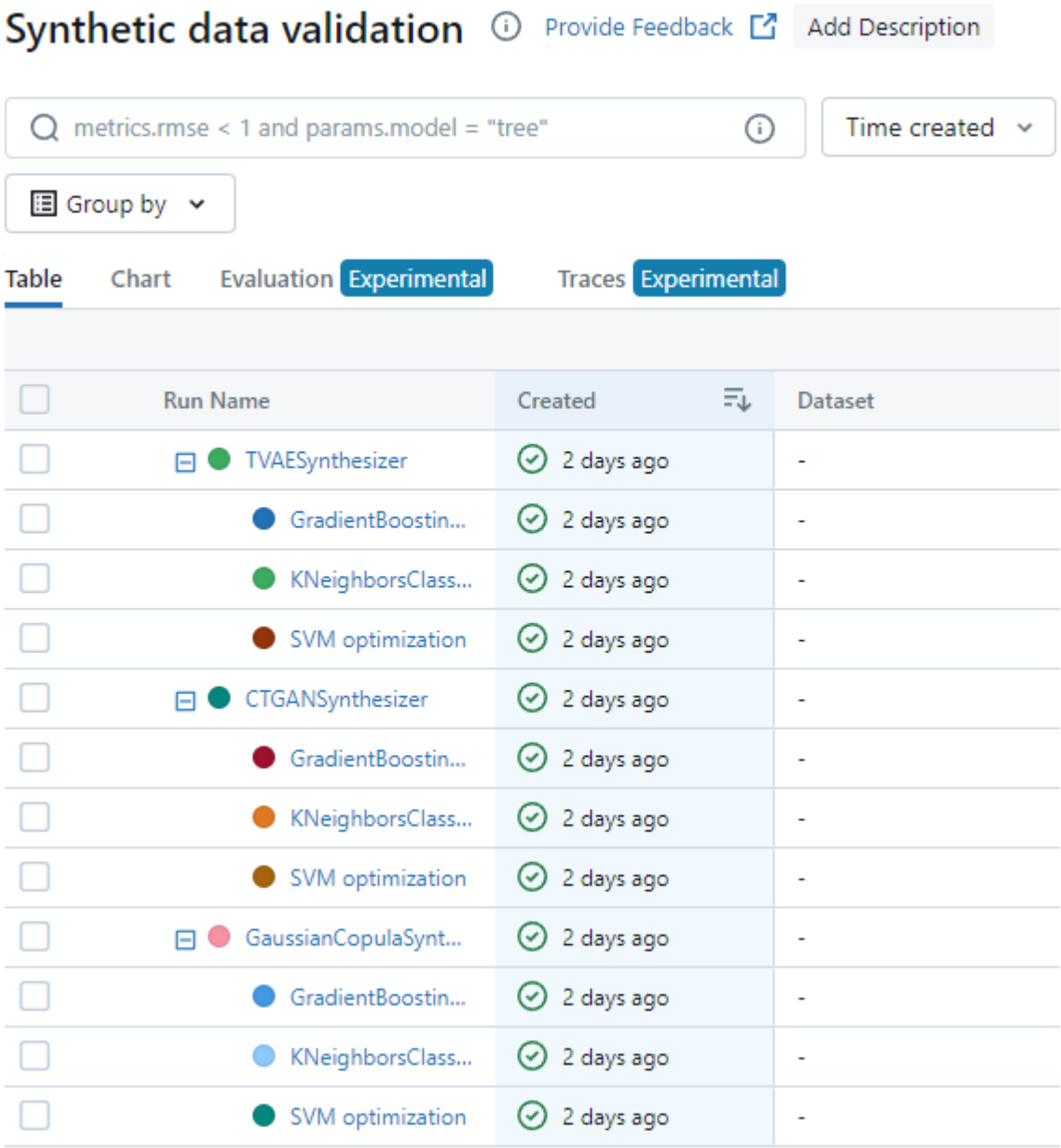
Логирование метрик качества работы алгоритмов, обученных на реальных и синтетических данных

GBC

SVC

KNeighbors

mlflow



Synthetic data validation ›

GradientBoostingClassifier optimization

Overview | Model metrics | System metrics | Artifacts

Description

No description

Details

Created at	2024-08-23 15:05:19
Created by	GANS
Experiment ID	759261123561527046
Status	Finished
Run ID	349e8d3ba7624b40a2af7d1181c765ec
Duration	3.1min
Parent run	CTGANSynthesizer
Datasets used	—
Tags	Add
Source	D:\Projects\neogen_venv\Lib\site-packages\ipykernel_launcher.py
Logged models	—
Registered models	—

Parameters (1)

Search parameters

Parameter	Value
synthetic_data_percent	0.9

Metrics (2)

Search metrics

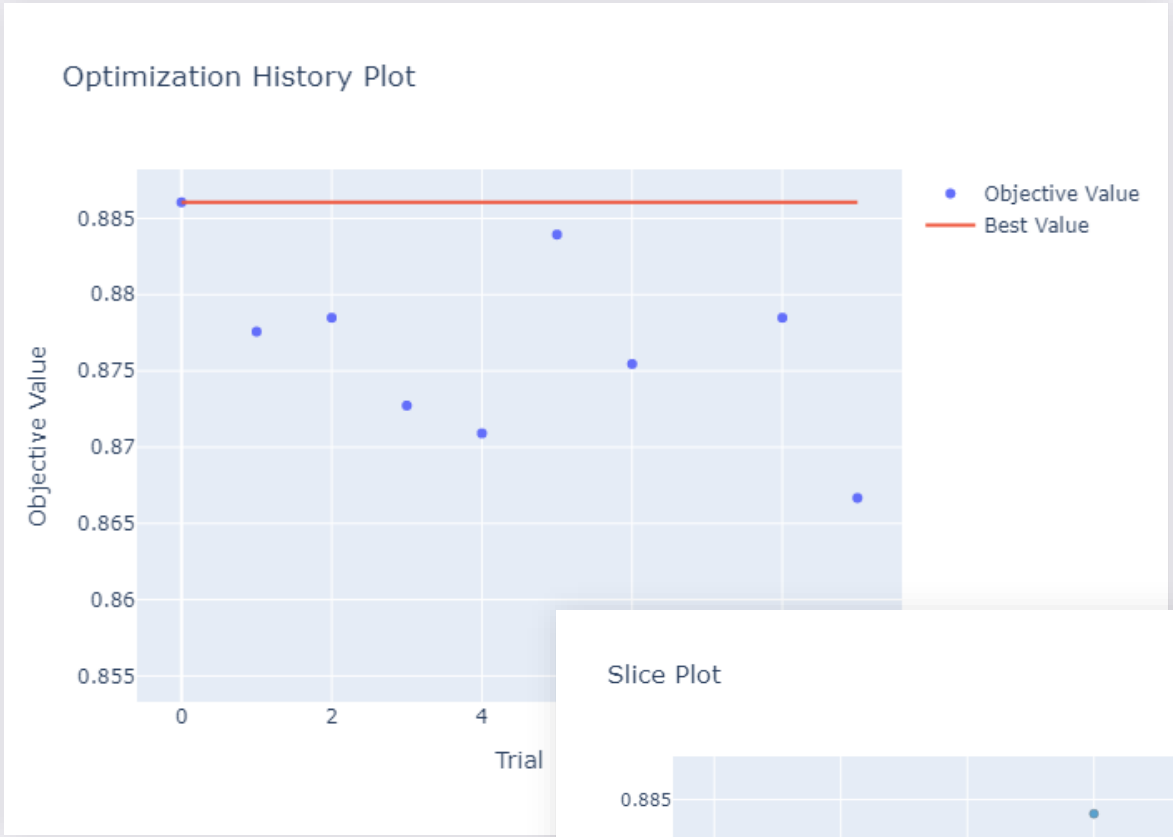
Metric	Value
real_accuracy	0.8142424242424242
synthetic_accuracy	0.8860606060606061

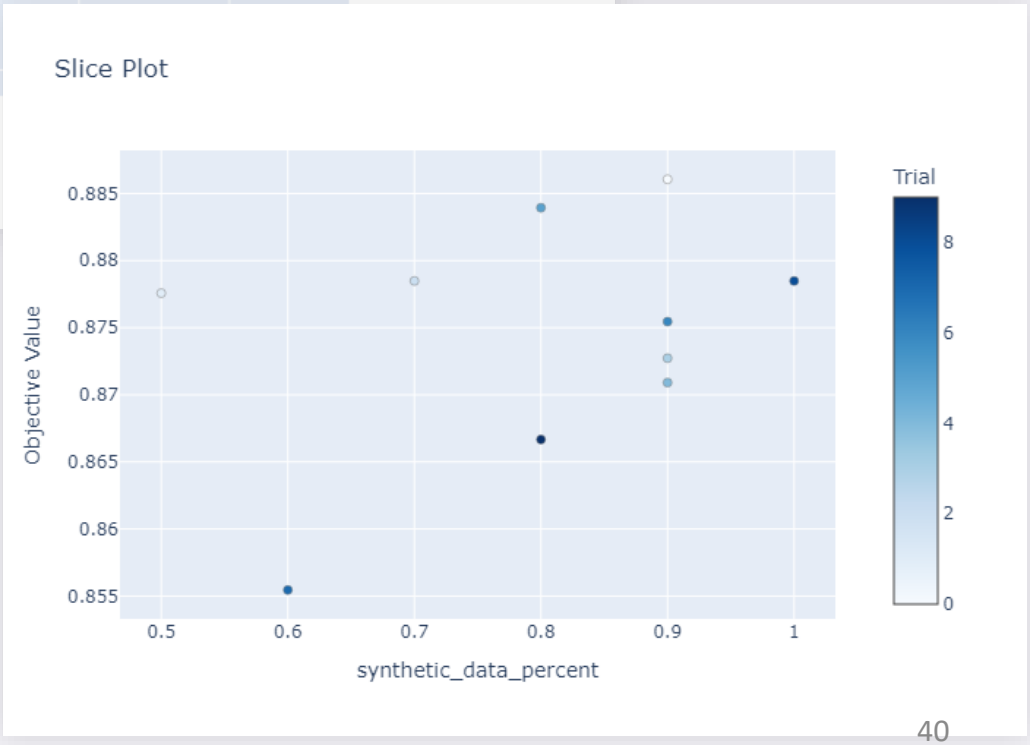
Synthetic data validation ›

GradientBoostingClassifier optimization

Overview | Model metrics | System metrics | Artifacts

- optimization
  - all_trials.csv
  - optimization_history.png
  - slice.png
- optimization_persent_0.5
  - all_trials.csv
  - optimization_history.png
  - slice.png
- optimization_persent_0.6
  - all_trials.csv
  - optimization_history.png
  - slice.png
- optimization_persent_0.7
  - all_trials.csv
  - optimization_history.png
  - slice.png
- optimization_persent_0.8
  - all_trials.csv
  - optimization_history.png
  - slice.png
- optimization_persent_0.9
  - all_trials.csv
  - optimization_history.png
  - slice.png
- optimization_persent_1.0
  - all_trials.csv
  - optimization_history.png
  - slice.png





Наше будущее

Улучшения фреймворка

To do

- Добавление возможности использования собственных ML моделей в пайплайн проверки синтетики
- Написание проверок дополнительных метрик качества синтетических датасетов (точности и конфиденциальности)
- Возможность генерации данных для отдельных классов целевой переменной в автоматическом режиме


Real Data
Synthetic Data

+7 (495) 984-25-13
info@neoflex.ru