

Хакатон {EESTech Challenge}

Решение команды SlovarikDB

Решенные нами задачи:

Задача

Выявление аномалий

Оценка доли аномалий в
наблюдениях

Оценка состояния трактора

Решение

Isolation Forest

Isolation Forest + временной
ряд по трактору

Использование “адаптивных”
моделей

Исходные данные

Набор данных: обработанная и “сырая” телеметрия по тракторам.

Разметка: метки аномалий и нормальных данных на части записей.

Ген. совокупность

Источники: anomaly +
normal + problems

Для чего?

Оценка
распределений,
Unsupervised learning

Оценка классов

Источники: normal +
problems

Для чего?

Валидация
классификатора,
обучение с учителем

Оценка трактора

Источники: сырые
данные

Для чего?

Оценка состояния
конкретного трактора
во времени

Очистка данных и инсайты

Проблема

Большое количество пустых колонок и “шума”

Огромное количество NaN значений

Разные виды распределений величин

Решение

Отбрасывание “мусорных” колонок и очистка шума

Заполнение NaN последними валидными данными

Использование “адаптивных” моделей

Реальные данные - не значит чистые данные. Но если их очистить - снизится и нагрузка на БД компании.

Классификация аномалий

Классические методы обучения с учителем не подходят для решения задачи, хоть у нас и есть метки классов.

Эксперимент: decision tree classifier

Древовидная модель после обучения на части данных обрабатывается на тестовой выборке. И работает феноменально. Но есть нюанс...

Метрики:

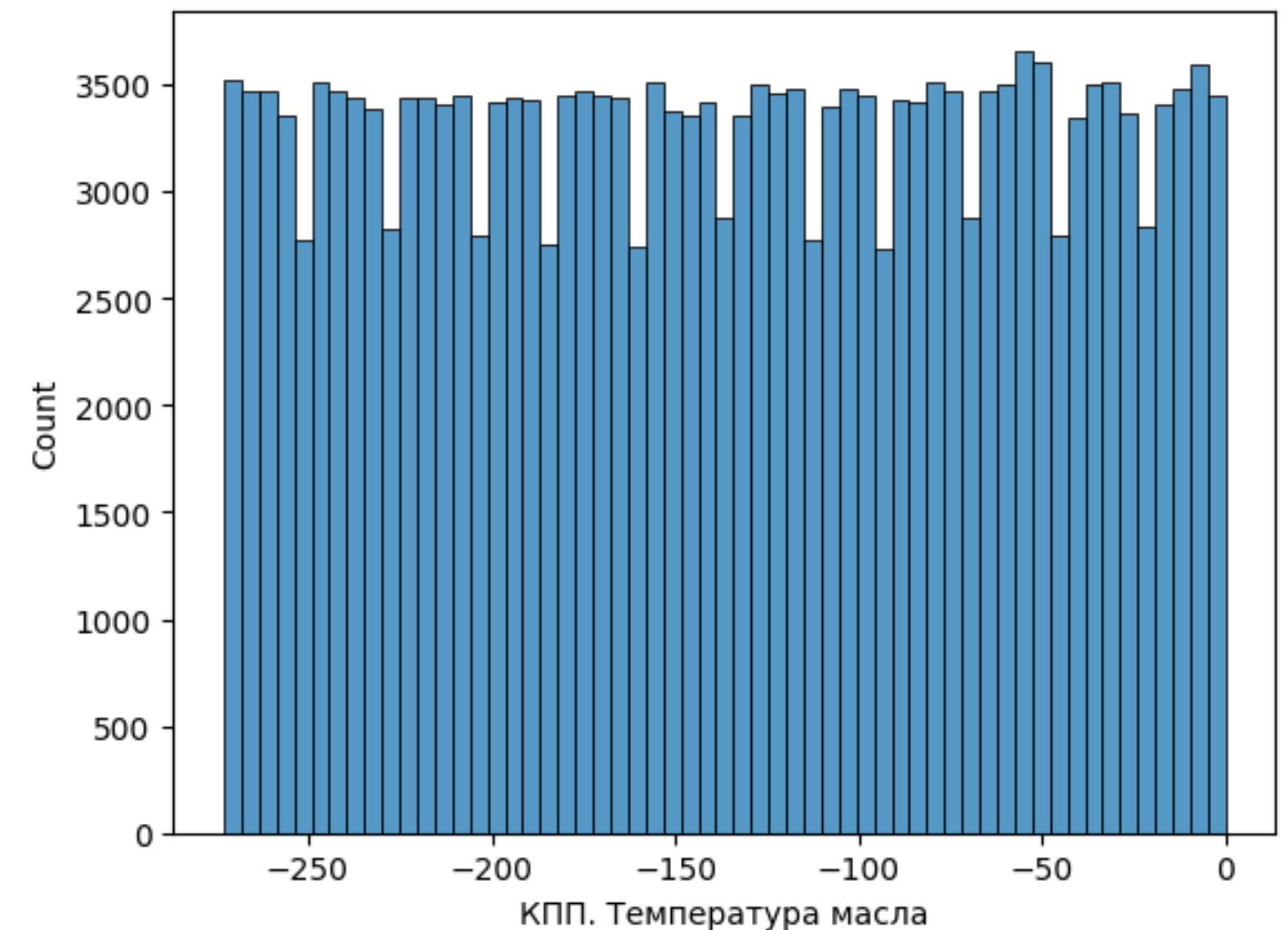
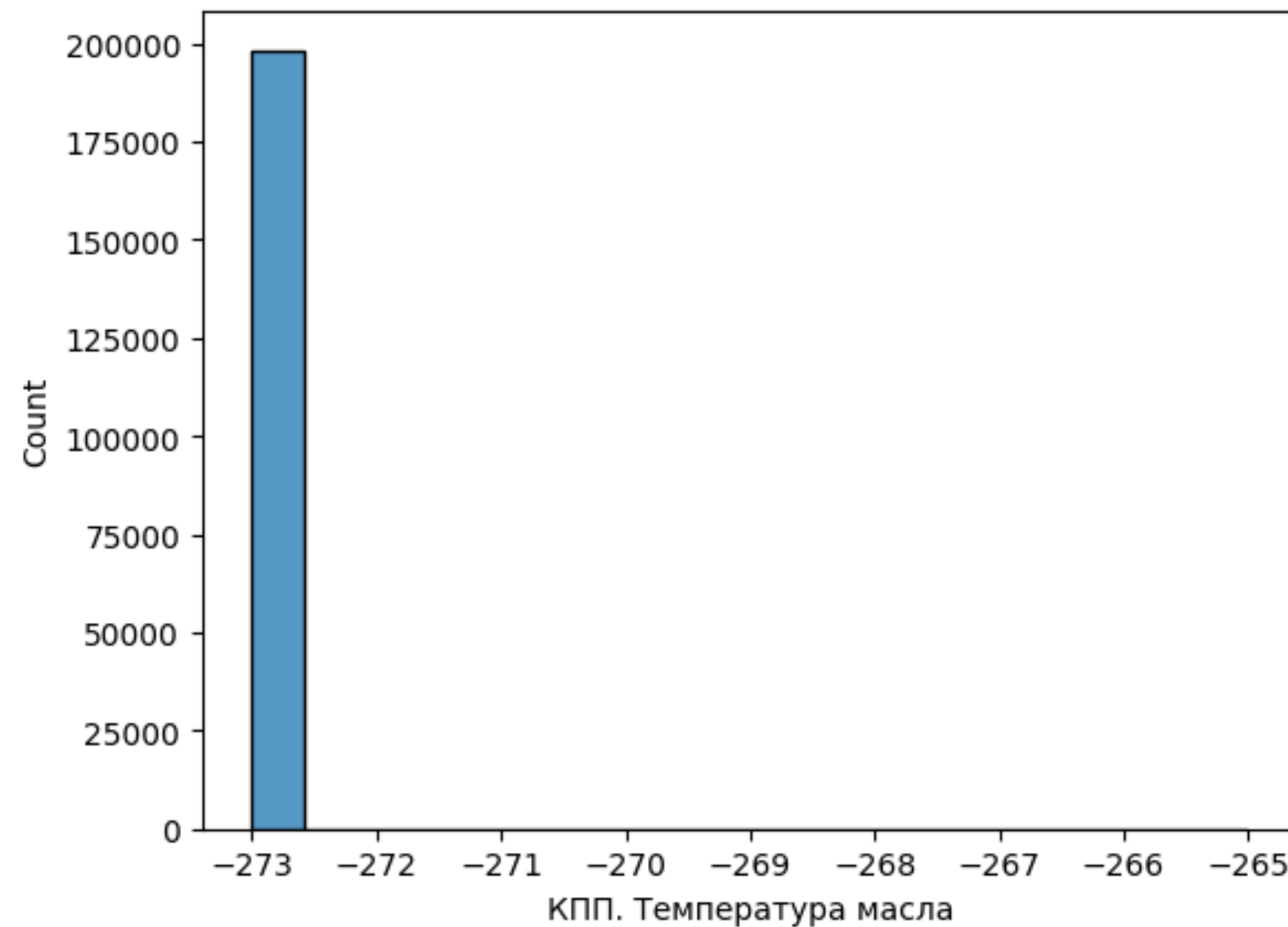
Precision: 1

Recall: 1

Acc: 1

Почему так?

Гипотеза: исходные метки “аномальности” были получены по отдельным экстремальным значениям.



Это может помешать масштабируемости подобной модели.

Альтернативный подход

Обучение без учителя + оптимизация гиперпараметров

Модель:

Isolation Forest:
выделяет аномалии
“подстраиваясь” под
распределение

Optuna:
Позволяет
оптимизировать
параметры IF

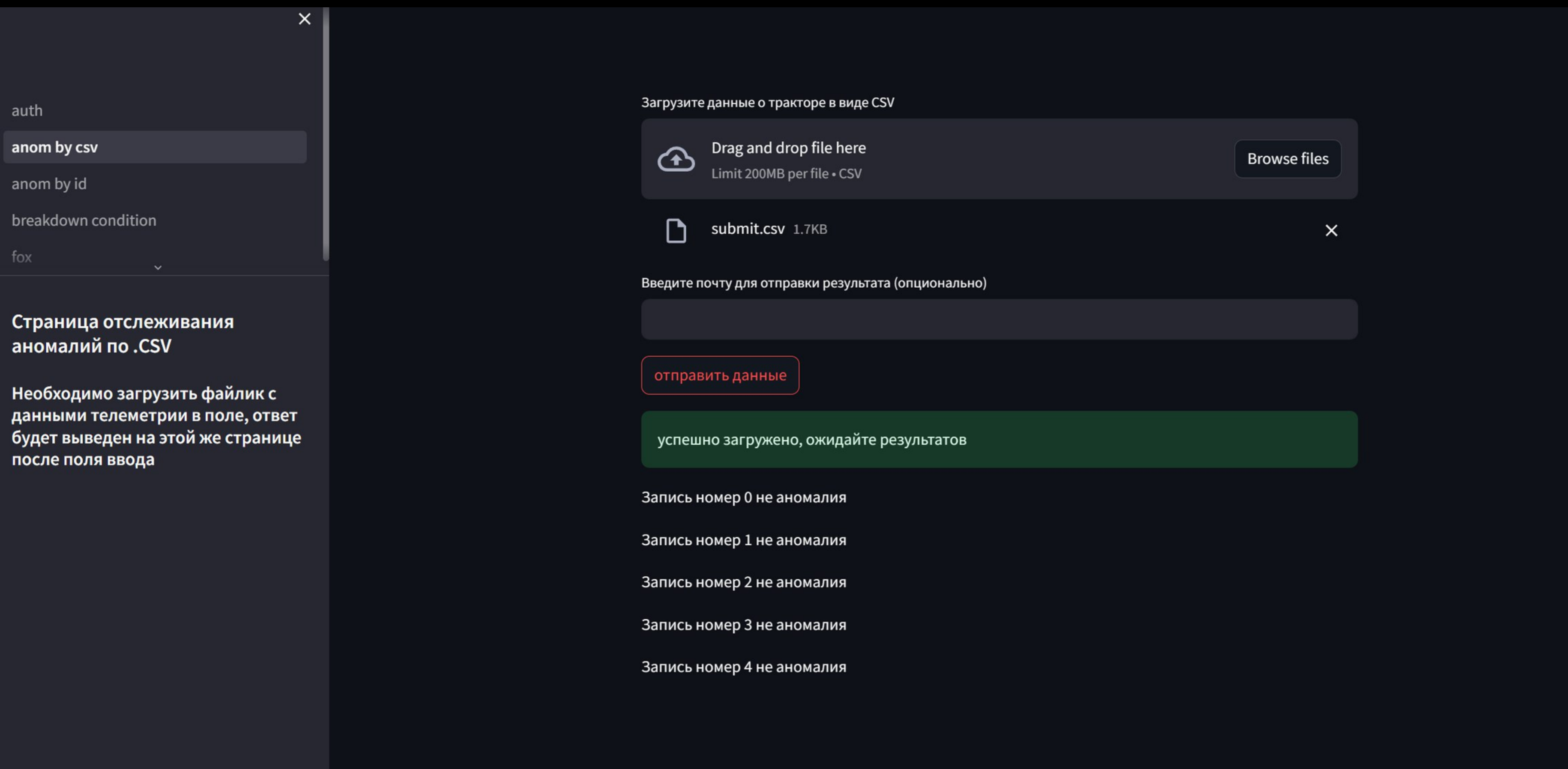
Метрики:
(по разметке)

Precision:
0.52

Recall:
1.0

Смысл? Данный подход выявляется в два раза больше аномалий, включая все в разметке.

Альтернативный подход UI



“Аномальный” трактор

Число и доля аномалий в наблюдениях конкретного трактора является важной метрикой.

User Case:

Диллер и владелец трактора могут видеть число аномалий для конкретного трактора. и их долю в общих наблюдениях

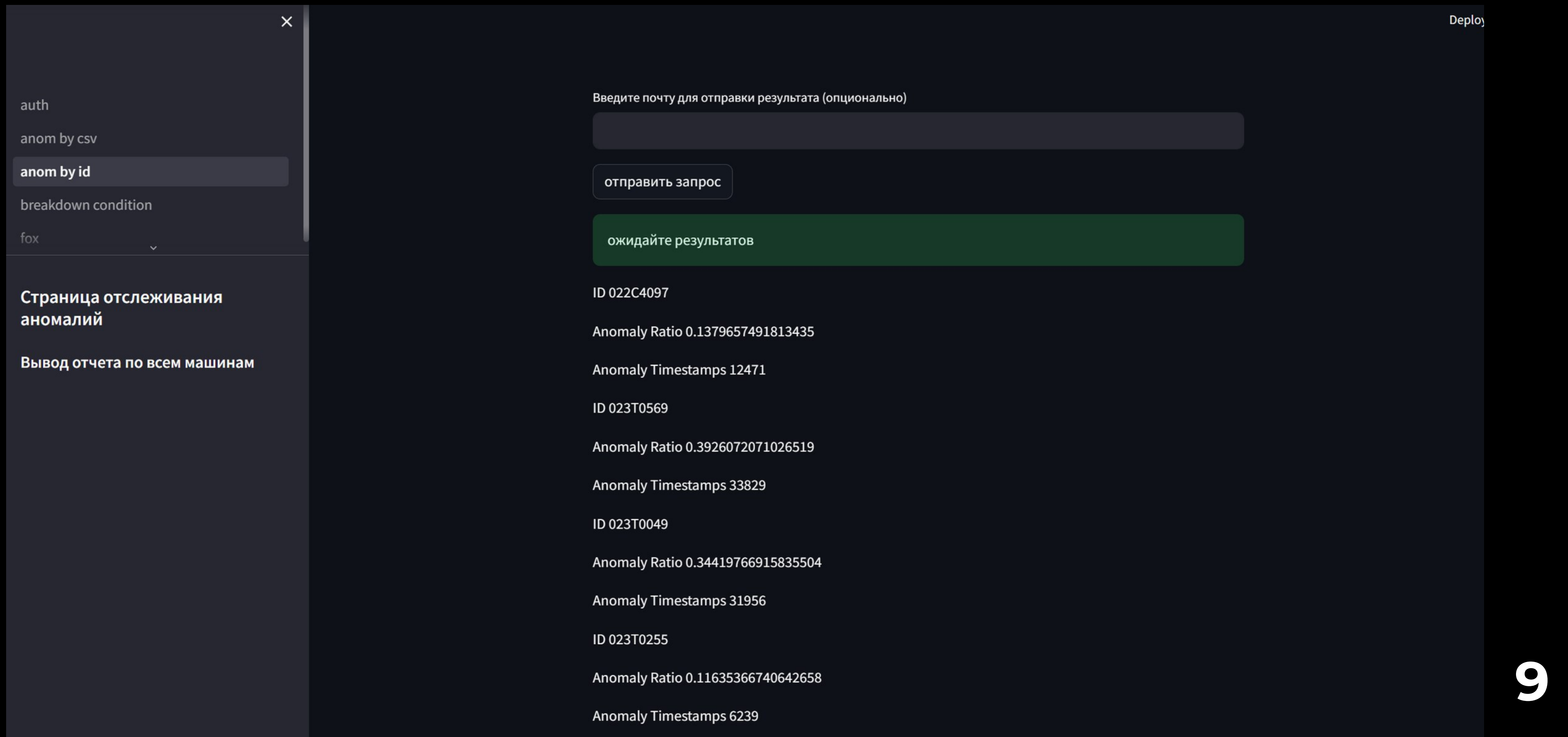
User Utility:

Анализ исправности трактора/
телеметрии

Realization:

Оценка значений временного
ряда через модель

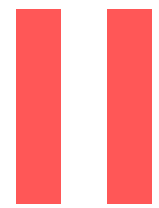
“Аномальный” трактор - UI



Оценка состояния трактора



Временные ряды по тракторам преобразуются в ряды по неделям и месяцам.



Значения текущего месяца и текущей недели оцениваются на то, являются ли они аномальными относительно ряда.



В случае, если текущие значения являются аномальными трактору рекомендуется пройти тех обслуживание.

Методика позволяет оценить какой конкретный параметр влияет на оценку, что позволяет привязать её к конкретной системе.

Оценка состояния трактора UI

auth

anom by csv

anom by id

breakdown condition

fox

Страница отслеживания состояний тракторов по ID (без пробелов)

Введите ID тракторов через запятую для получения информации об их состоянии

Введите ID тракторов

C1969,P3106

Введите почту для отправки результата (опционально)

отправить данные

ожидайте результатов

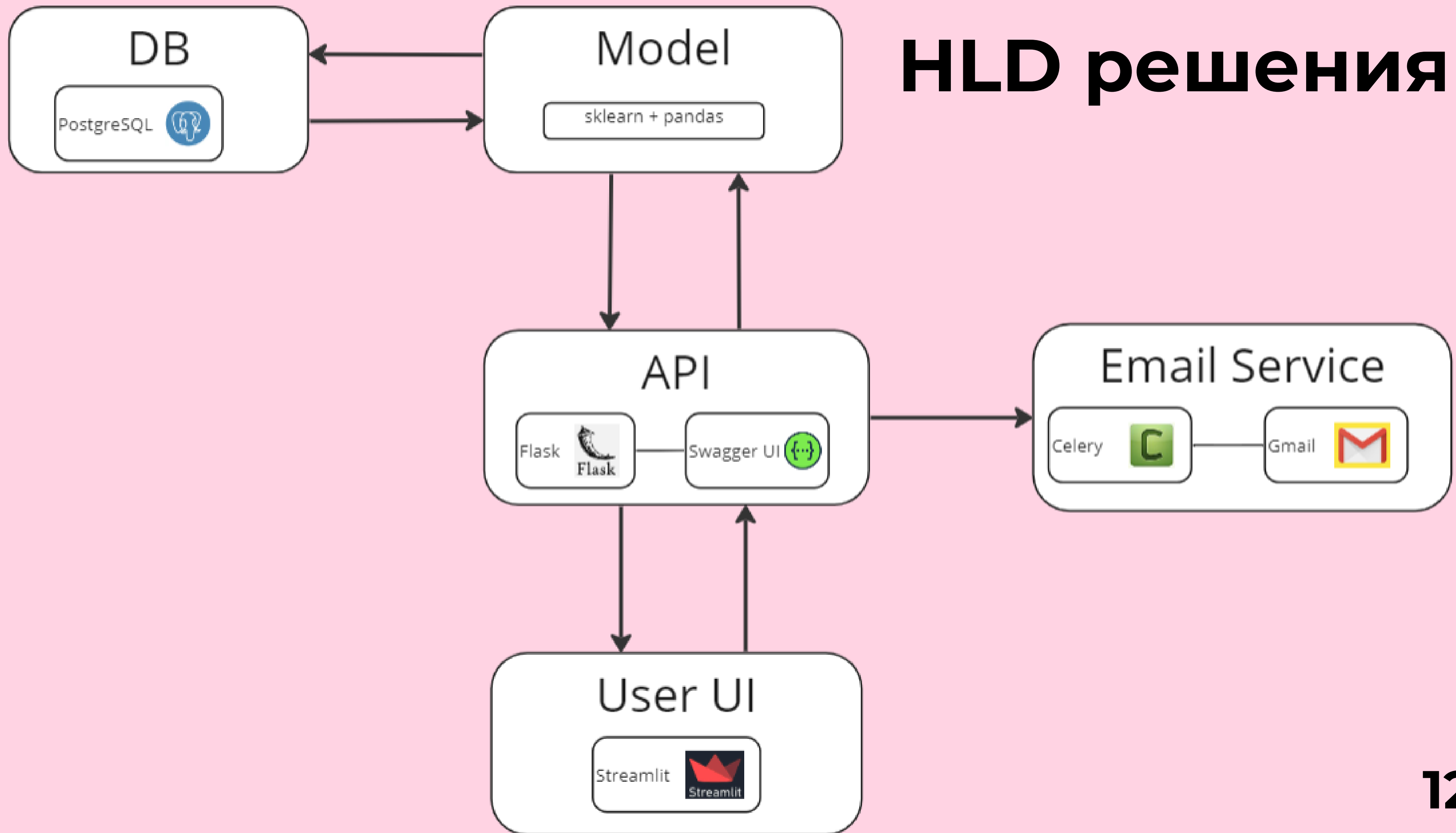
ID C1969 : По данным за последний месяц машина рекомендуется к проверке

ID P3106 : По данным за последний

месяц

 машина в порядке

HLD решения



Заключение и инсайты

Далеко не всегда самые сложные решения - лучшие.

Система телеметрии нуждается в улучшении с точки зрения ограничения передачи мусорных данных

Наше решение легко масштабируемо и внедряемо , так как не требует постоянной разметки данных.

СПАСИБО ЗА ВНИМАНИЕ!

*С вами были Богодист Всеволод,
Артем Вичук и Сахабутдинов
Рустам*