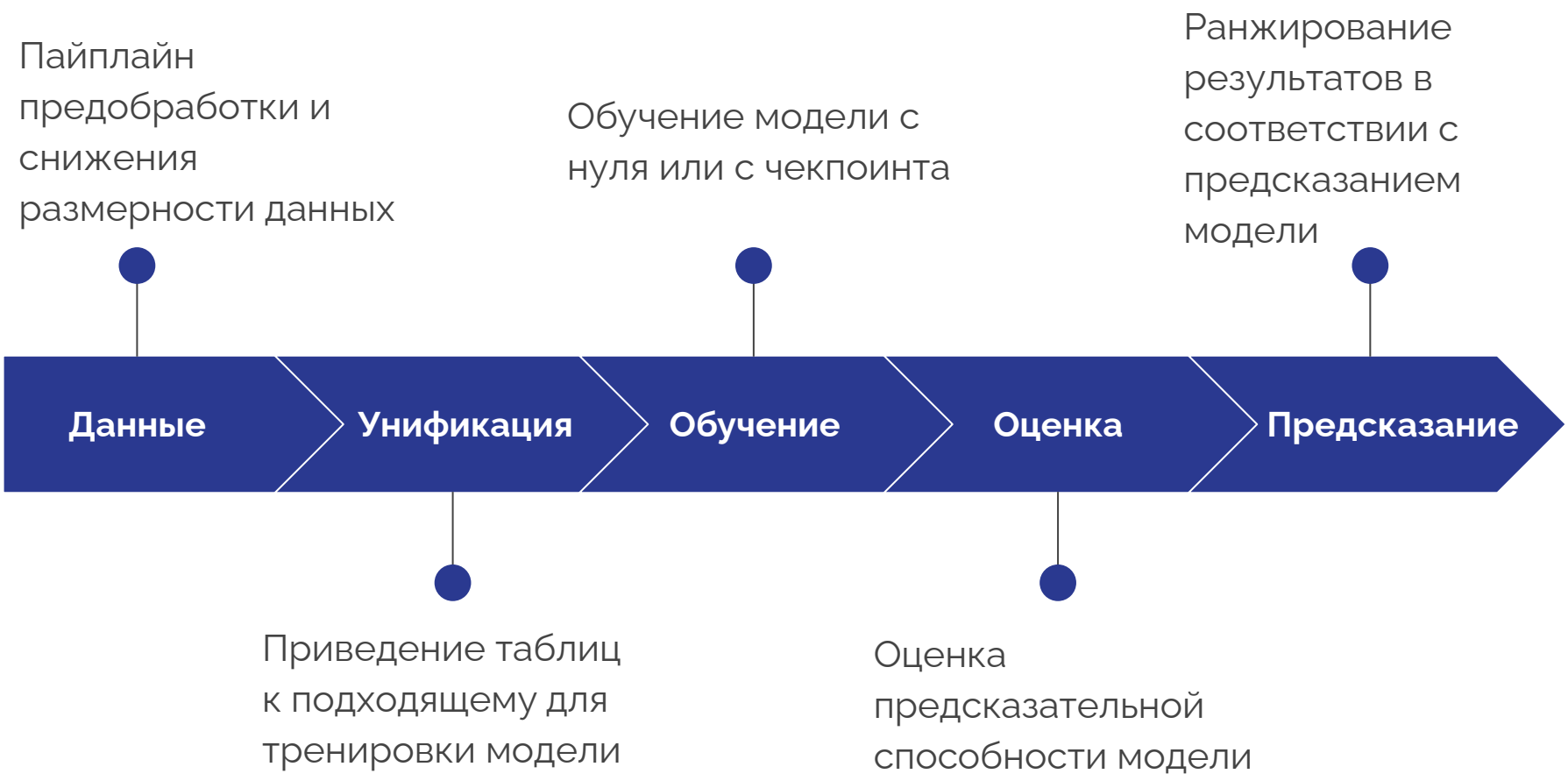
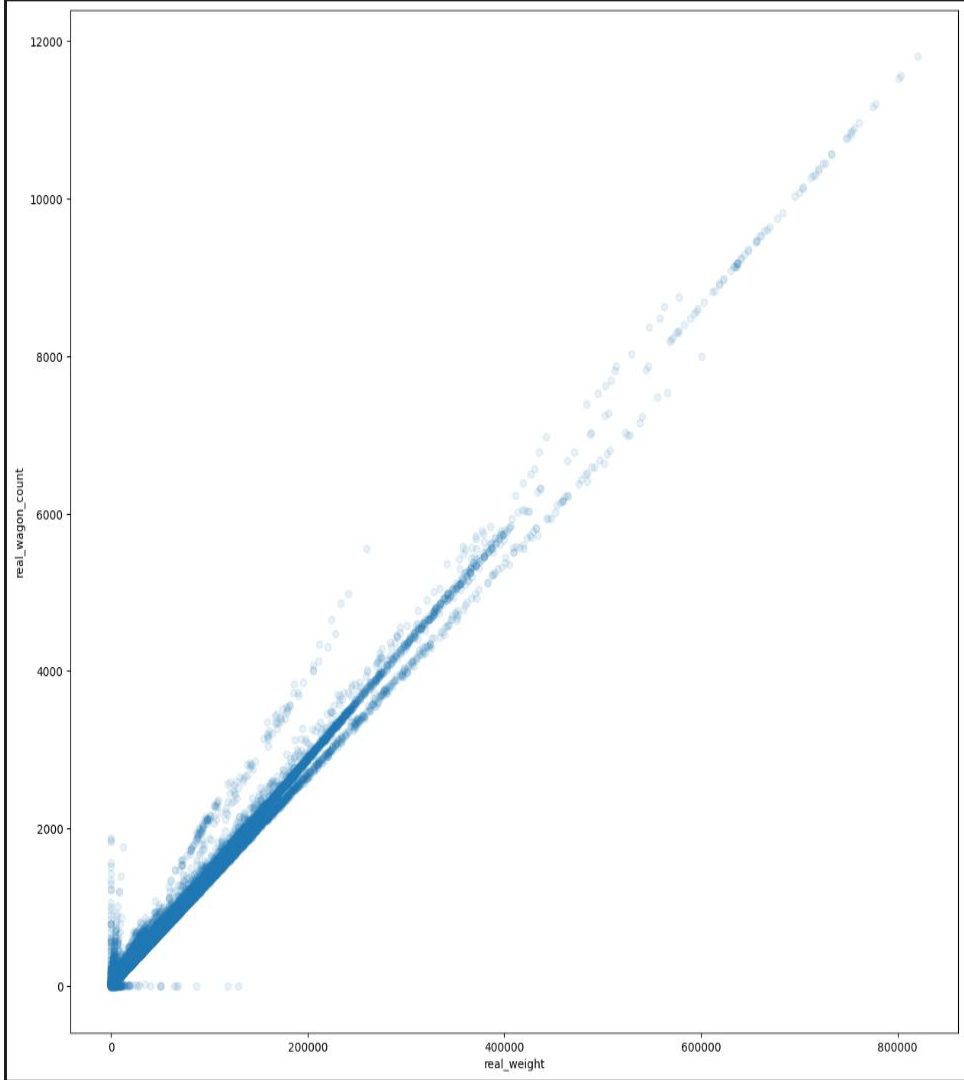
Решение команды RDX

Data Wagon трек 1


Пайплайн предобработки и снижения размерности данных
Обучение модели с нуля или с чекпоинта
Ранжирование результатов в соответствии с предсказанием модели
Данные
Унификация
Обучение
Оценка
Предсказание
Приведение таблиц к подходящему для тренировки модели виду
Оценка предсказательной способности модели

Выбросы данных

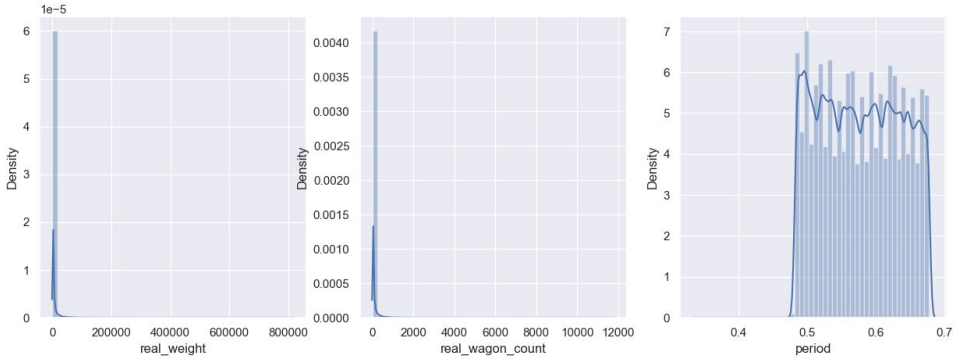
Фильтрация



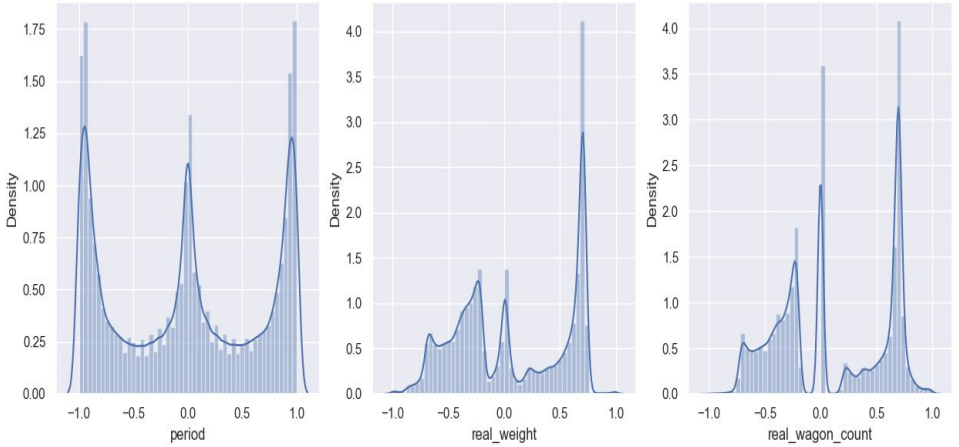
Нормализация

Скейлинг

До нормализации

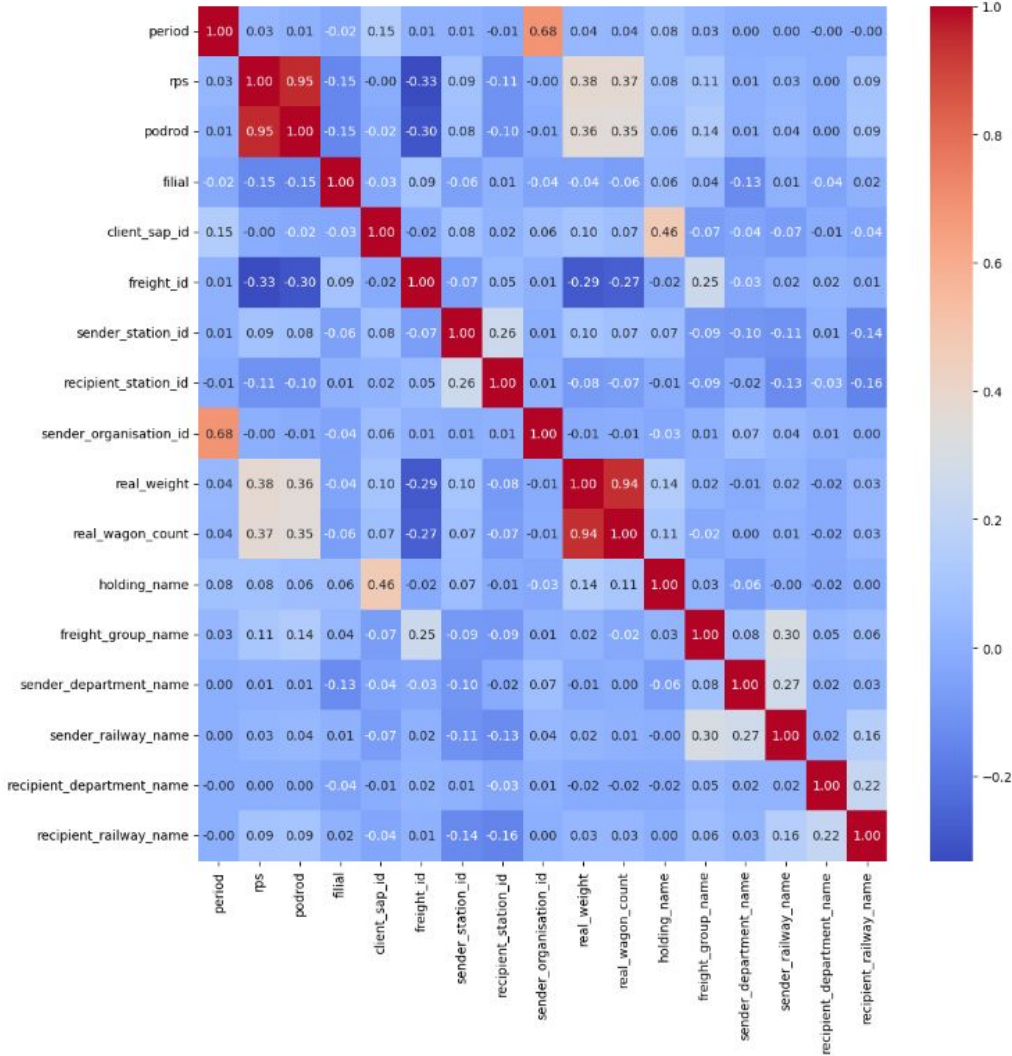


После нормализации и скейлинга



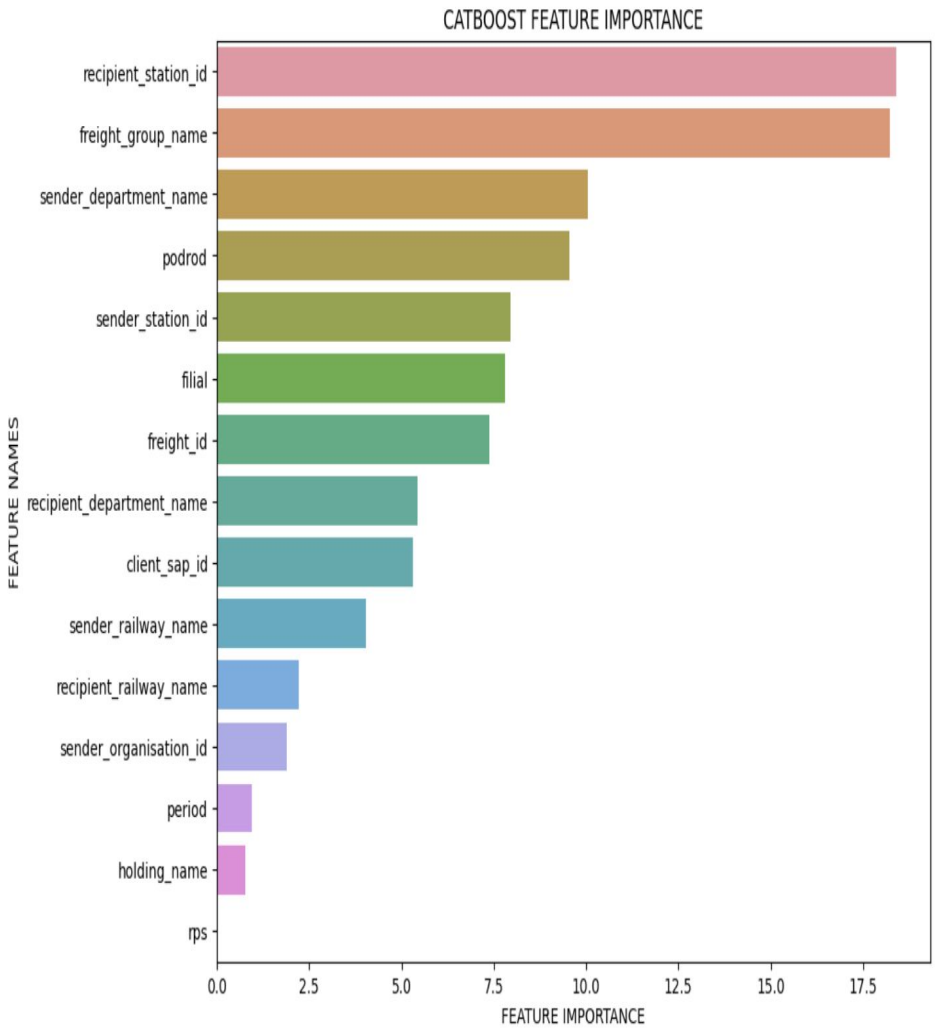
Обработка данных

Матрица корреляций фичей



Влияние фичей

На результат предсказания

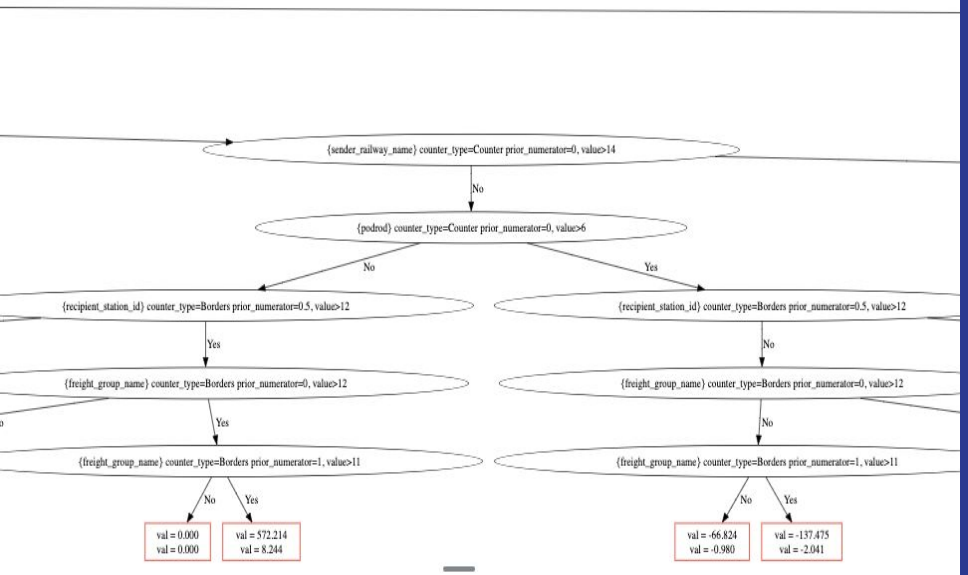


Решение

CatBoost

Градиентный бустинг деревьев решений с пайплайном предобработки данных, валидацией модели и дообучением с чекпоинта

Часть структуры дерева



Подобранные оптимальные параметры

Depth: 10
Iterations: 190
Learning Rate: 0.2

Возможные доработки

Данные

Перенос обработки данных в *Airflow* и унификация процесса обработки данных, а также дальнейшего сохранения в *feature store*

Модель

Вынос работы модели в отдельный сервис, при помощи *MLFlow* и настройка регулярного дообучения и обновления артефактов на основе *S3*

Оценка

Внедрение карты метрик и динамический сбор пользовательского фидбека с последующей обработкой и передачей в модель