BigQuery-Geotab Intersection Congestion

Александр Гридин Екатерина Вековцева Марк Колпаков Родион Гудзь

Содержание

Состав команды

Проект

BigQuery-Geotab Intersection Congestion

Состав команды

- **Александр Гридин**
- **Екатерина Вековцева**
- **Марк Колпаков**
- **Родион Гудзь**

Данные

Источник

Датасет **BigQuery-Geotab Intersection Congestion** с соревнования на **Kaggle**

Описание

Агрегированные показатели регистрации поездок коммерческих транспортных средств, таких как грузовики.

Группировка

Сгруппированы по перекрестку, месяцу, времени суток, направлению движения через перекресток и наличию выходного дня.

Цель проекта

Бизнес-контекст

Оптимизация логистики коммерческого транспорта через анализ времени простоя на перекрёстках

Целевые переменные:

- `TotalTimeStopped_p20` - 20-й перцентиль
- `TotalTimeStopped_p50` - медиана
- `TotalTimeStopped_p80` - 80-й перцентиль

Метрика качества:

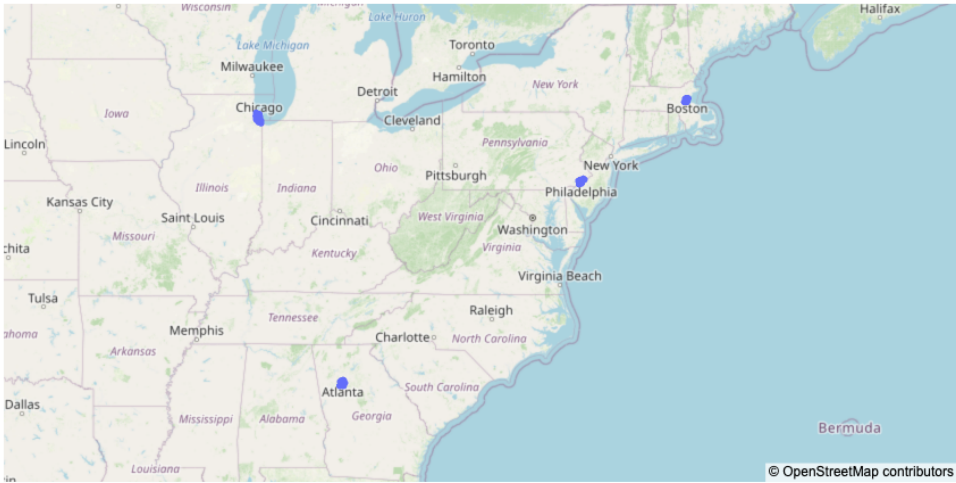
- Основная: MAE (Mean Absolute Error)
- Мультитаргетная оценка

EDA — Предобработка целей

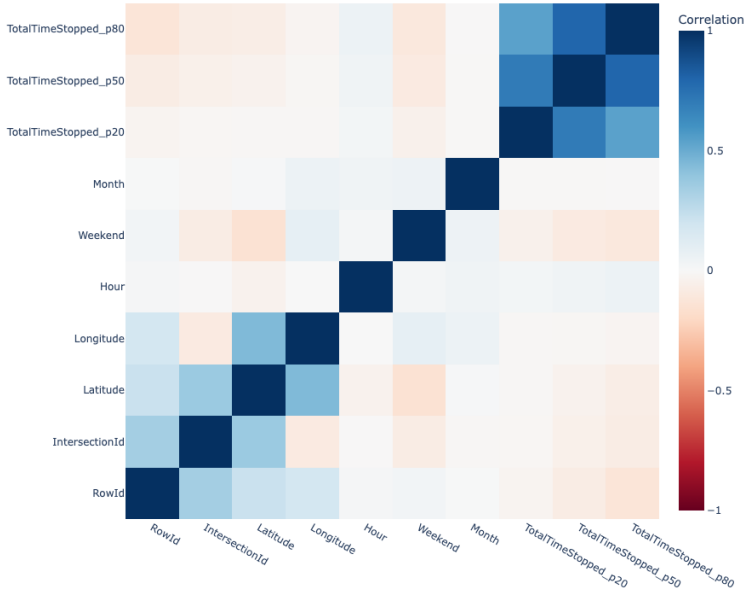
Действие

Убраны лишние целевые переменные, чтобы не допустить утечки информации

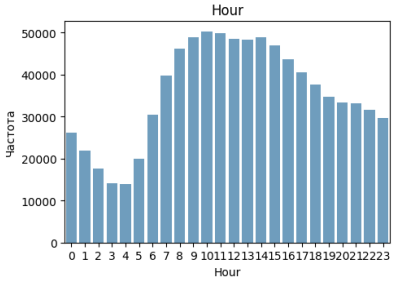
Intersections with unknown streets names

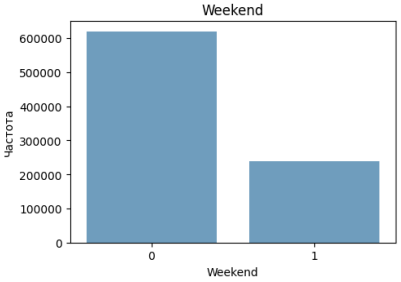


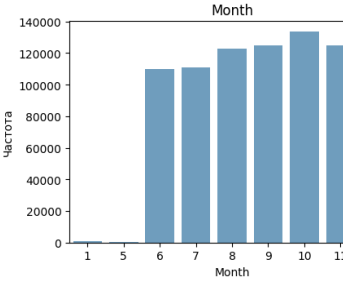
Матрица корреляции признаков



Train dataset







Анализ выбросов и клиппинг

Методы

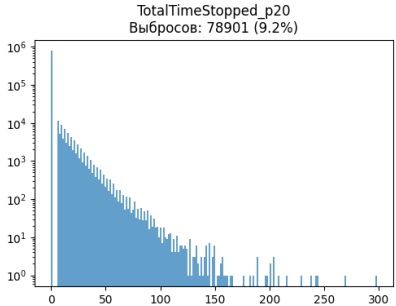
Проведен анализ выбросов с помощью **IQR**, **Z-оценки**, **теста Граббса** и выполнен клиппинг по **квантилю 99,5%** для целевой переменной

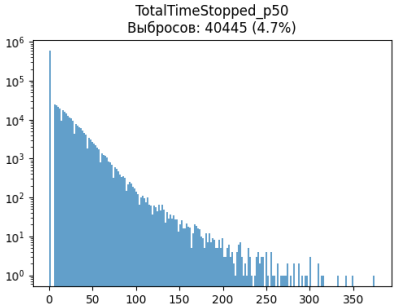
Наблюдение

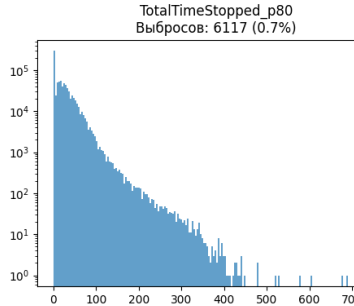
Выявлено отсутствие выбросов в координатах и **ID**

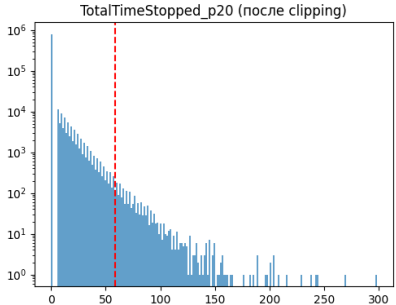
Дополнительно

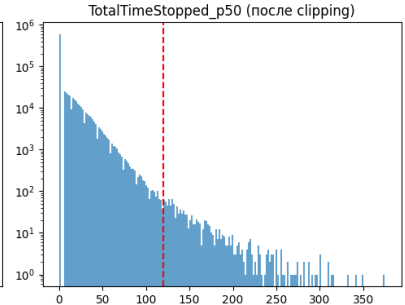
Добавлены флаги **редкого перекрёстка**, **редких входных и выходных улиц**, **редких путей**

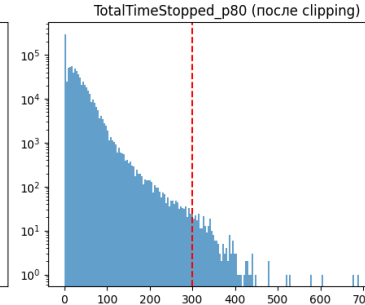

TotalTimeStopped_p20
Выбросов: 78901 (9.2%)
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
50
100
150
200
250
300


TotalTimeStopped_p50
Выбросов: 40445 (4.7%)
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
50
100
150
200
250
300
350


TotalTimeStopped_p80
Выбросов: 6117 (0.7%)
10^5
10^4
10^3
10^2
10^1
10^0
0
100
200
300
400
500
600
70


TotalTimeStopped_p20 (после clipping)
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
50
100
150
200
250
300


TotalTimeStopped_p50 (после clipping)
10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
50
100
150
200
250
300
350


TotalTimeStopped_p80 (после clipping)
10^5
10^4
10^3
10^2
10^1
10^0
0
100
200
300
400
500
600
70

Анализ выбросов и клиппинг — таблица порогов

Клиппинг аномалий - стратегия

Пороги клиппинга:

Переменная	99.5% квантиль	Порог
p20	298 сек	60 сек
p50	375 сек	120 сек
p80	763 сек	300 сек

Эффект клиппинга на распределении

Результат

Удаление менее 0.5% экстремальных значений

Логика порогов:

- Физическая реалистичность
- Сохранение 99.5% данных
- Учёт бизнес-логики

Методы обнаружения сложных аномалий

Isolation Forest:

- Contamination: 1%
- Находит глобальные аномалии
- Быстрая работа на больших данных

Результат сравнения

Только 149 общих аномалий из 4000 найденных

Кодирование категорий и соседние признаки

Кодирование

Target Encoding с кросс-валидацией и **Label Encoding** для категориальных признаков

Качество после кодирования

После кодирования MAE: **p20 = 2.859844**, **p50 = 9.184705**, **p80 = 17.539677**

Новые пространственные признаки

+3 признака, основанных на ближайших соседях:

- среднее значение таргета у k ближайших соседей
- расстояние до ближайшего соседа

Генерация новых признаков - категориальные

Target Encoding для City:

- K-Fold схема (k=5)
- Smoothing = 20
- Отдельно для каждого таргета

Вывод

Существенные различия между городами - важный признак

Label Encoding для:

- EntryStreetName
- ExitStreetName
- Path
- EntryHeading
- ExitHeading

Генерация новых признаков - геопространственные

На основе ближайших соседей:

- Среднее таргета у 10 соседей
- Расстояние до ближайшего соседа
- Плотность в радиусе 500м

Преимущество

Помогает с новыми перекрёстками (cold-start)

Метрика расстояния:

- Евклидово расстояние
- Перевод в метры (111 км/градус)

Контекстные признаки

Доменные знания в признаках

1. **Часы пик:** Расстояние до 8:00 и 17:00
2. **Сложные повороты:** Флаг левого поворота/разворота
3. **Редкие маршруты:** Частота Path < 50
4. **Выходные в больших городах:** Комбинация Weekend и City

Распределение углов поворота

Влияние часов пик

Обзор данных

Объём данных:

- Train: 856,000 строк
- Test: 1,800,000 строк
- 30 признаков в исходных данных

Ключевые признаки:

- Координаты (Latitude, Longitude)
- Время (Hour, Month, Weekend)
- География (City, Street names)
- Направления движения

Особенности тестовой выборки:

- На 11% больше уникальных перекрёстков в Test
- Данные разделены в случайном порядке
- Требует учёта cold-start проблемы

Вывод

Необходимы методы, устойчивые к новым перекрёсткам

Географическое распределение перекрёстков

Расположение перекрёстков в США (выборка 100k точек)

- **Наблюдение:** Перекрёстки сосредоточены в крупных городах
- **Вывод:** Важно учитывать региональные особенности
- **Решение:** Добавление признаков на основе кластеризации координат

Распределение данных по месяцам и часам

Распределение по месяцам

- Данные с июня по декабрь
- Нет годовой цикличности

Распределение по часам

- Равномерное покрытие всех часов
- Часы пик: 8:00 и 17:00

Важный вывод

Цикличность по месяцам отсутствует, но есть суточные паттерны

Пропущенные значения и их анализ

Тепловая карта пропущенных значений (missingno)

Пропуски в:

- EntryStreetName
- ExitStreetName

Причина:

- Съезды с магистралей
- Частные территории

Решение

Пропуски не случайны - создаём отдельный признак "Unknown street"

Архитектура модели CatBoost

Параметры модели:

- Задача: MultiRMSE (3 таргета)
- Итерации: 200-600
- Глубина: 5-6
- Learning rate: 0.05-0.1
- Early stopping: 30-50 раундов

Категориальные признаки:

- Автоматическая обработка
- Оптимальное кодирование

Кривая обучения (train/validation)

Находка

CatBoost лучше ручного кодирования категорий

Динамика качества на разных этапах

Этап	p20 MAE	p50 MAE	p80 MAE	Изменение
Начальный бейзлайн	2.860	9.185	17.540	–
+ Обработка аномалий	2.810	9.050	16.990	-1.7%
+ Новые признаки	2.840	9.120	17.210	+0.8%
+ Отбор признаков	2.850	9.150	17.250	+0.3%

Таблица: Сравнение качества на валидации (подвыборка 200k)

Ключевые выводы

- Обработка аномалий дала наибольший прирост
- Ручная генерация признаков не всегда улучшает качество
- CatBoost хорошо справляется с raw-признаками

Сравнение моделей

Модели

Сравнили 2 разных класса моделей: **линейную Ridge** и **ансамблевую RandomForest**.

Результат

По MAE лучше оказалась линейная модель: **Ridge MAE = 9.37**, **RF MAE = 10.28**.

Сравнение моделей: Ridge vs Random Forest

Ridge Regression

SHAP важность признаков (Ridge)

- MAE: 9.37
- Интерпретируемость: Высокая

Random Forest

SHAP важность признаков (RF)

- MAE: 10.28
- Интерпретируемость: Средняя

Выбор для интерпретации

Ridge показал лучшее качество и более стабильные SHAP значения

Топ-признаки по влиянию (SHAP анализ)

Общие для обеих моделей:

1. City_te_p80 (кодированный город)
2. Weekend (выходной день)
3. hour_dist_to_peak
4. Latitude, Longitude
5. is_rare_path

Наблюдение

17 общих признаков в топ-20 у Ridge и RF

Сравнение важности признаков

- Зелёный: Одинаковый знак влияния
- Красный: Разный знак влияния

Локальная интерпретация: LIME объяснения

LIME объяснение для Ridge

- Преобладают OHE признаки улиц
- Чёткие линейные эффекты

LIME объяснение для Random Forest

- Более сложные взаимодействия
- Нелинейные эффекты

Вывод по методам интерпретации

SHAP лучше для глобальной интерпретации, LIME - для локальной

Анализ SHAP-эмбеддингов и аномалий

SHAP-эмбеддинги:

- Вектора вкладов каждого признака
- Размерность: число признаков
- Анализ распределений

Обнаружение аномалий:

- 4% аномальных профилей
- Isolation Forest + Кластеризация
- Удаление ухудшает качество

Важный результат

Удаление SHAP-аномалий увеличило MAE с 7.20 до 7.30

Характеристики кластеров SHAP-эмбеддингов

Кластер	Размер	Средний таргет	Топ признаки
Cluster 0	24%	6.31 сек	Hour, Weekend, координаты
Cluster 1	18%	8.45 сек	EntryStreetName, is_rare_path
Cluster 2	15%	12.67 сек	City_te, hour_dist_to_peak
Cluster 3	22%	18.92 сек	Все признаки равномерно
Cluster 4	12%	25.34 сек	Редкие улицы, сложные повороты
Cluster 5	9%	32.42 сек	Экстремальные значения всех признаков

Таблица: Характеристики кластеров SHAP-эмбеддингов

Практическое применение

- Кластер 0: Стандартные условия - малые задержки
- Кластер 5: Экстремальные условия - большие задержки
- Добавление cluster как признака дало MAE 7.23 (vs 7.20 baseline)

Эксперимент: SHAP-эмбеддинги как новые признаки

Идея эксперимента:

- Использовать SHAP-эмбеддинги как фичи
- Обучить линейную модель на них
- Сравнить с оригинальными признаками

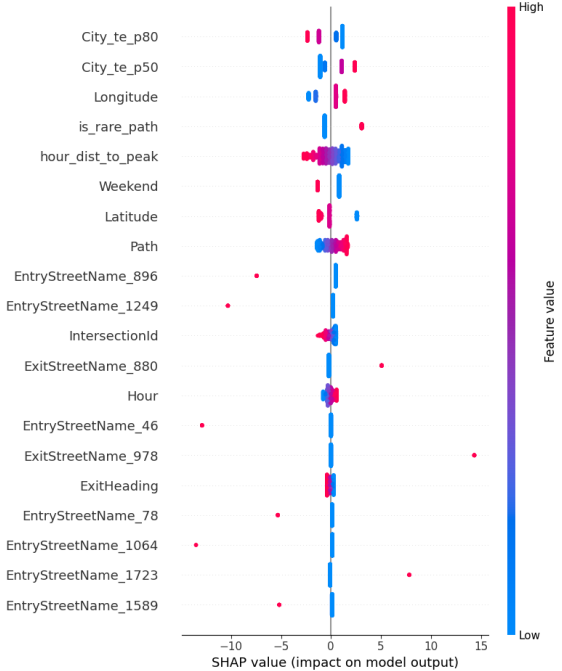
Результаты (3-fold CV):

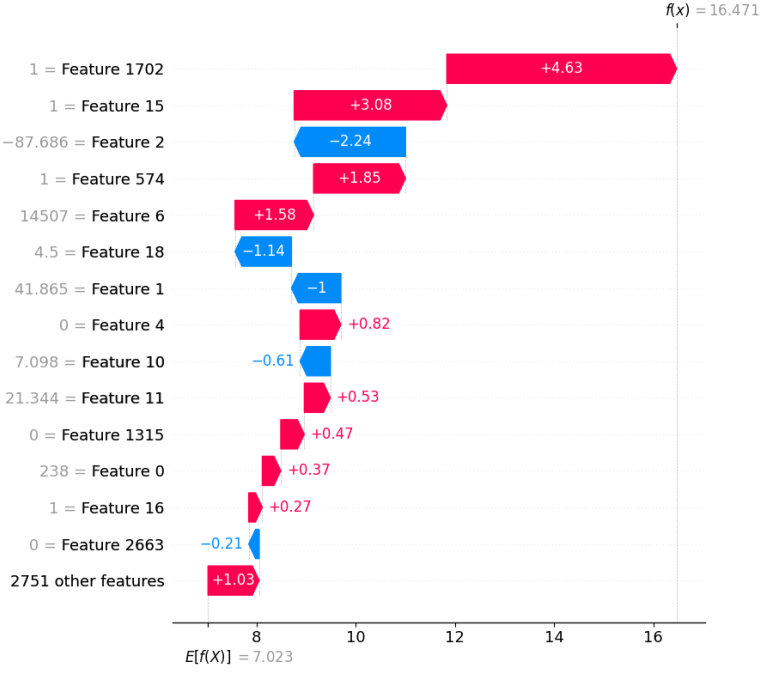
- CatBoost (оригинал): 6.79 MAE
- Ridge (только SHAP): 9.05 MAE
- Ridge (оригинал+SHAP): 9.37 MAE

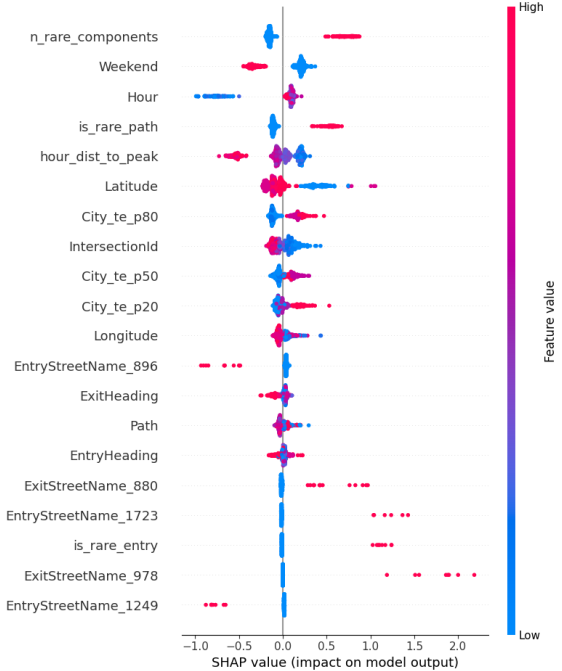
Сравнение качества моделей

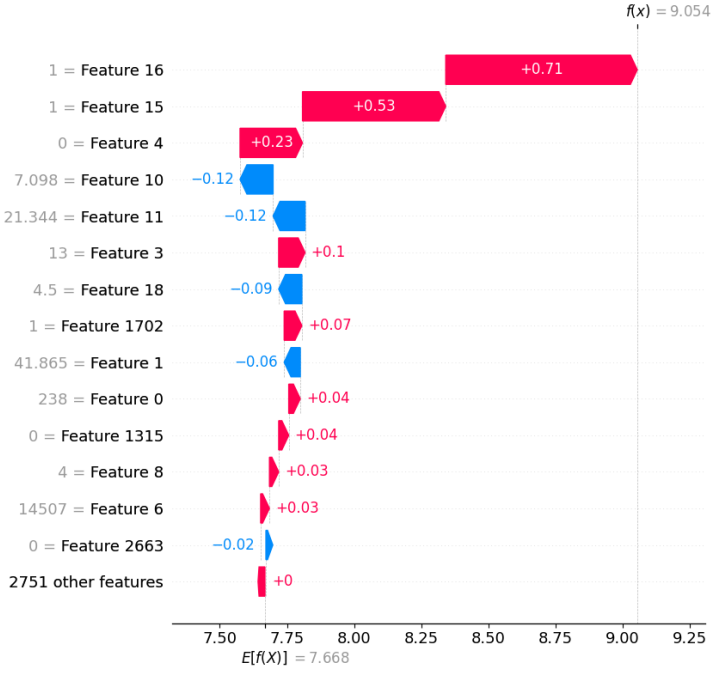
Вывод

SHAP-эмбеддинги - хороший инструмент диагностики, но не замена оригинальным признакам


City_te_p80
City_te_p50
Longitude
is_rare_path
hour_dist_to_peak
Weekend
Latitude
Path
EntryStreetName_896
EntryStreetName_1249
IntersectionId
ExitStreetName_880
Hour
EntryStreetName_46
ExitStreetName_978
ExitHeading
EntryStreetName_78
EntryStreetName_1064
EntryStreetName_1723
EntryStreetName_1589
−10
−5
0
5
10
15
SHAP value (impact on model output)
High
Feature value
Low


f(x) = 16.471
1 = Feature 1702
+4.63
1 = Feature 15
+3.08
−87.686 = Feature 2
−2.24
1 = Feature 574
+1.85
14507 = Feature 6
+1.58
4.5 = Feature 18
−1.14
41.865 = Feature 1
−1
0 = Feature 4
+0.82
7.098 = Feature 10
−0.61
21.344 = Feature 11
+0.53
0 = Feature 1315
+0.47
238 = Feature 0
+0.37
1 = Feature 16
+0.27
0 = Feature 2663
−0.21
2751 other features
+1.03
8
10
12
14
16
E[f(X)] = 7.023


n_rare_components
Weekend
Hour
is_rare_path
hour_dist_to_peak
Latitude
City_te_p80
IntersectionId
City_te_p50
City_te_p20
Longitude
EntryStreetName_896
ExitHeading
Path
EntryHeading
ExitStreetName_880
EntryStreetName_1723
is_rare_entry
ExitStreetName_978
EntryStreetName_1249
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
SHAP value (impact on model output)
High
Feature value
Low


f(x) = 9.054
1 = Feature 16
+0.71
1 = Feature 15
+0.53
0 = Feature 4
+0.23
7.098 = Feature 10
−0.12
21.344 = Feature 11
−0.12
13 = Feature 3
+0.1
4.5 = Feature 18
−0.09
1 = Feature 1702
+0.07
41.865 = Feature 1
−0.06
238 = Feature 0
+0.04
0 = Feature 1315
+0.04
4 = Feature 8
+0.03
14507 = Feature 6
+0.03
0 = Feature 2663
−0.02
2751 other features
+0
7.50
7.75
8.00
8.25
8.50
8.75
9.00
9.25
E[f(X)] = 7.668

Выводы по SHAP и кластеризации

Удаление атипичных SHAP-профилей

Очистка наблюдений с атипичным SHAP-профилем ухудшила качество (**MAE 7.204 → 7.305**), поэтому удаление по данному критерию нецелесообразно.

Кластеризация SHAP-эмбеддингов

Кластеризация выявила осмысленную сегментацию (лучший силуэт у **agglomerative**), однако добавление признака **cluster** дало лишь умеренный эффект и не превзошло бейзлайн (**MAE_with_cluster = 7.234**).

SHAP + исходные признаки

Добавление исходных признаков к SHAP-эмбеддингам в линейной модели не улучшило качество, а ухудшило его примерно на **0.32 MAE**. В обоих вариантах качество существенно хуже, чем у исходной модели **CatBoost (MAE = 6.79 ± 0.31)**.

Ключевые результаты проекта

Успехи:

- Обработка аномалий дала +1.7% качества
- CatBoost показал лучшие результаты
- SHAP анализ выявил важные факторы
- Созданы осмысленные кластеры данных

Вызовы:

- Ручные признаки не всегда улучшают качество
- Новые перекрёстки (cold-start)
- Мультитаргетная природа задачи

Финальные метрики

- Лучшая модель: CatBoost с обработкой аномалий
- MAE p50: 9.05 (улучшение на 1.4%)

Дальнейшие направления работы

1. **Временные ряды:** Учёт последовательности наблюдений
2. **Графовые нейросети:** Моделирование структуры дорог
3. **Ансамбли:** Комбинация разных типов моделей
4. **Online learning:** Адаптация к новым данным

Потенциальный impact

- Снижение времени доставки на 3-5%
- Оптимизация маршрутов в реальном времени
- Прогнозирование заторов на перекрёстках

Вопросы и ответы

Спасибо за внимание!

Контакты:
Команда проекта "BigQuery-Geotab Intersection Congestion"
Kaggle Competition