# Полный технический отчет

## Проект: "Движок Геоповеденческого Профилирования"

Версия: 2.0

Дата: 14.09.2025

Автор: Самат Кабдыгали

Ссылка на гугл коллаб

<https://colab.research.google.com/drive/1JXL5pxkfFadjFq8lNED6tsra51FCex1X?usp=sharing>

<https://colab.research.google.com/drive/1WTweh3aFdLYWGeHvPjRZHCXZVnxEV6ec?usp=sharing>

### Содержание

1. **Введение и стратегическое видение (Стр. 2)**
   * 1.1. Резюме для руководства
   * 1.2. Проблема: ограничения традиционного анализа
   * 1.3. Наше видение: от сервиса такси к Mobility Intelligence
2. **Фаза 1: Подготовка данных и инжиниринг признаков (Стр. 3)**
   * 2.1. Анализ исходных данных: структура и ограничения
   * 2.2. Процедура очистки и предобработки
     + *Техническая реализация и Output*
   * 2.3. Научный сдвиг: от временных к пространственным производным
     + *Техническая реализация и Output*
   * 2.4. Агрегация признаков: формирование "ДНК" поездки
     + *Техническая реализация и Output*
   * 2.5. Результат Фазы 1: итоговый датасет признаков
3. **Фаза 2: Поведенческая кластеризация и поиск аномалий (Стр. 6)**
   * 3.1. Цель: неконтролируемая сегментация стилей вождения
   * 3.2. Подход "Baseline-First": снижение размерности с помощью PCA
     + *Техническая реализация и Output*
   * 3.3. Кластеризация методом K-Means: выявление поведенческих архетипов
     + *Техническая реализация и Output*
   * 3.4. Обнаружение аномалий с помощью Isolation Forest
     + *Техническая реализация и Output*
   * 3.5. Результат Фазы 2: обогащенный поведенческий профиль поездки
4. **Фаза 3: Гео-анализ и создание карты рисков (Стр. 9)**
   * 4.1. Цель: переход от анализа "кто" к анализу "где"
   * 4.2. Геопространственная индексация H3: технология и обоснование выбора
   * 4.3. Методология расчета risk\_score
     + *Техническая реализация и Output*
   * 4.4. Методология расчета demand\_score
     + *Техническая реализация и Output*
   * 4.5. Результат Фазы 3: геоповеденческий слой данных
5. **Фаза 4: Прототипирование и "RL-Фасад" (Стр. 12)**
   * 5.1. Цель: визуализация и демонстрация бизнес-применения
   * 5.2. Технологический стек прототипа
   * 5.3. Компонент 1: Интерактивная карта рисков и спроса
     + *Output для пользователя*
   * 5.4. Компонент 2: Анализ "ДНК" водителя
     + *Output для пользователя*
   * 5.5. Инновация "RL-Фасад": имитация системы принятия решений
     + *Техническая реализация и Output*
   * 5.6. Результат Фазы 4: работающий интерактивный прототип
6. **Заключение и дальнейшее развитие (Стр. 15)**
   * 6.1. Итоги и достигнутые результаты
   * 6.2. Бизнес-ценность и стратегическое применение
   * 6.3. Направления для дальнейшего развития

### 1. Введение и стратегическое видение

#### 1.1. Резюме для руководства

Этот проект представляет собой создание аналитической системы нового поколения, которая решает две стратегические задачи для inDrive:

1. **Проактивное повышение безопасности:** Переход от реагирования на инциденты к их предотвращению путем выявления потенциально рискованных стилей вождения и опасных участков городской среды.
2. **Создание уникального дата-актива:** Формирование "геоповеденческой карты" города, которая является ценным ресурсом для оптимизации операций, повышения эффективности платформы и открывает новые возможности для монетизации данных.

Наше решение анализирует анонимные геотреки, создавая поведенческий профиль ("ДНК") каждой поездки и, на его основе, строит динамическую карту городских рисков и спроса. Это не просто аналитический отчет, а фундамент для более умной, безопасной и эффективной платформы.

#### 1.2. Проблема: ограничения традиционного анализа

Стандартные подходы к анализу телематических данных часто сталкиваются с проблемой неполноты и нерегулярности данных. В нашем случае, отсутствие точных временных меток и переменная частота отправки гео-точек делают классические методы, основанные на временных производных (например, ускорение как dV/dt), нестабильными и ненадежными. Такой анализ не позволяет объективно сравнивать поездки между собой и формировать целостную картину.

#### 1.3. Наше видение: от сервиса такси к Mobility Intelligence

Мы предлагаем изменить парадигму: рассматривать анонимные геотреки не как побочный продукт операционной деятельности, а как ценнейший сырьевой ресурс. Цель проекта — построить движок, который превращает этот ресурс в стратегический актив — **"Геоповеденческую сигнатуру города"**. Этот актив позволит inDrive перейти от роли сервиса по перевозке к роли компании, обладающей глубоким пониманием транспортной динамики города (Mobility Intelligence), что открывает новые горизонты для оптимизации и монетизации.

### 2. Фаза 1: Подготовка данных и инжиниринг признаков

На этом этапе была заложена научная и техническая основа всего проекта. Цель — превратить сырые, зашумленные и нерегулярные данные в структурированный и информативный датасет, пригодный для машинного обучения.

#### 2.1. Анализ исходных данных: структура и ограничения

* **Источник:** Исходный файл geotracks.csv.
* **Структура:** Каждая строка представляет собой одну гео-точку со следующими полями:
  + lat, lng: Географические координаты.
  + spd: Скорость в данной точке.
  + azm: Азимут (направление движения).
  + randomized\_id: Уникальный анонимный идентификатор поездки.
* **Ключевые ограничения:**
  1. **Отсутствие временных меток (timestamp):** Невозможно напрямую рассчитать ускорение и другие временные характеристики.
  2. **Нерегулярная частота точек:** Расстояние между точками варьируется, что делает простые разностные метрики несопоставимыми.
  3. **Наличие "шума":** Присутствие статичных точек (например, во время ожидания пассажира) и коротких, неинформативных поездок.

#### 2.2. Процедура очистки и предобработки

Для обеспечения качества данных был реализован многоступенчатый процесс очистки.

##### Техническая реализация и Output

* 1. **Удаление статичных точек:** Мы отфильтровали все точки со скоростью (spd) менее 1 км/ч. Это необходимо, чтобы исключить из анализа периоды простоя (ожидание на парковке, долгие светофоры), так как они не характеризуют стиль вождения в движении и вносят шум в данные.
  2. **Фильтрация коротких треков:** Поездки, состоящие менее чем из 10 точек, были удалены. Такие поездки слишком коротки, чтобы на их основе можно было сделать статистически значимые выводы о поведении водителя.
  3. **Обработка аномальных значений:** Был проведен анализ на наличие неправдоподобных значений скорости и азимута (например, скорость > 200 км/ч в городских условиях).

**Простыми словами об Output:** На выходе этого шага мы получили файл geotracks\_cleaned.parquet. Представьте, что мы "просеяли" исходные данные через сито. Мы выбросили весь "мусор": поездки из двух-трех точек, которые ни о чем не говорят, и долгие стоянки на месте. В итоге осталась только чистая, полезная информация — точки, которые отражают реальное движение автомобилей.

#### 2.3. Научный сдвиг: от временных к пространственным производным

Это ключевая инновация проекта, позволившая обойти проблему отсутствия временных меток.

* **Проблема:** Классическое ускорение (a=ΔtΔV​) рассчитать невозможно.
* **Наше решение:** Мы перешли от производных по времени к **производным по расстоянию**. Это позволило измерять интенсивность маневров относительно пройденного пути, делая метрики инвариантными к пропускам во времени.

**Введенные метрики:**

1. **Изменение скорости на метр (speed\_change\_per\_m):** Показывает, насколько интенсивно водитель ускоряется или тормозит на данном участке пути. Высокие значения характерны для стиля "газ-тормоз".
2. **Интенсивность поворота на метр (turn\_rate\_per\_m):** Отражает, насколько резко водитель маневрирует. Высокие значения соответствуют частым и резким перестроениям.

##### Техническая реализация и Output Для каждой поездки мы сдвигали данные на одну строчку, чтобы в каждой строке была информация о текущей (p\_i) и следующей (p\_{i+1}) точке. Затем попарно рассчитывали расстояние (haversine\_vector) и применяли формулы: $ \text{speed\_change\_per\_m} = \frac{|\text{spd}{i+1} - \text{spd}i|}{\text{distance}(p\_i, p{i+1})} $ $ \text{turn\_rate\_per\_m} = \frac{\min(|\text{azm}{i+1} - \text{azm}i|, 360 - |\text{azm}{i+1} - \text{azm}i|)}{\text{distance}(p\_i, p{i+1})} $Простыми словами об Output: Результатом этого шага стал тот же датафрейм, но с двумя новыми, очень важными столбцами: speed\_change\_per\_m и turn\_rate\_per\_m. Теперь, глядя на любую точку маршрута, мы можем сказать не просто "здесь машина ехала со скоростью 60 км/ч", а "именно в этом месте водитель очень резко нажал на газ/тормоз" или "здесь был совершен очень резкий поворот". Мы превратили сырые координаты в осмысленное поведение.

#### 2.4. Агрегация признаков: формирование "ДНК" поездки

Чтобы перейти от анализа отдельных точек к профилю всей поездки, мы агрегировали вычисленные пространственные метрики для каждого randomized\_id.

##### Техническая реализация и Output Мы сгруппировали датафрейм по randomized\_id и для каждой группы (поездки) рассчитали несколько статистик (mean, std, p95) для наших новых признаков.

* + **Среднее (mean):** Отражает типичное поведение водителя.
  + **Стандартное отклонение (std):** Показывает стабильность стиля вождения.
  + **95-й перцентиль (p95):** Характеризует склонность к экстремальным маневрам.

**Простыми словами об Output:** Представьте, что каждая поездка — это длинный музыкальный трек. Агрегация — это как сделать краткое саммари этого трека. Мы получили новую, гораздо более компактную таблицу, где каждая строка — это уже не точка, а целая поездка. А столбцы — это ее характеристики: "в среднем водитель ездит спокойно", "но иногда склонен к очень резким маневрам", "его стиль вождения очень нервный и нестабильный". Эта таблица и есть "ДНК" каждой поездки.

#### 2.5. Результат Фазы 1: итоговый датасет признаков

По итогам фазы был создан датасет dna\_features\_pca.csv, где каждая строка представляет одну уникальную поездку, а столбцы — ее поведенческие характеристики ("ДНК"). Этот датасет является полностью подготовленным для применения моделей машинного обучения.

### 3. Фаза 2: Поведенческая кластеризация и поиск аномалий

Цель этого этапа — автоматически, без предварительной разметки, сегментировать все поездки на группы со схожим стилем вождения и выявить наиболее нетипичные, аномальные треки.

#### 3.1. Цель: неконтролируемая сегментация стилей вождения

Мы применили методы обучения без учителя (unsupervised learning) для выявления естественных кластеров в данных. Это позволяет системе самой находить существующие "архетипы" водителей.

#### 3.2. Подход "Baseline-First": снижение размерности с помощью PCA

* **Проблема "проклятия размерности":** Датасет с "ДНК" имел большое количество признаков. Работа с таким пространством напрямую неэффективна.
* **Решение:** Мы применили метод главных компонент (PCA).

##### Техническая реализация и Output

* 1. **Масштабирование:** Все признаки были отмасштабированы с помощью StandardScaler. Это как привести все валюты к доллару, чтобы их можно было сравнивать.
  2. **Применение PCA:** Метод позволил сжать исходное многомерное пространство признаков в 2 главные компоненты, сохранив при этом максимальное количество информации.

**Простыми словами об Output:** Представьте, что "ДНК" водителя — это 10 разных ползунков (скорость, резкость и т.д.). PCA помог нам заменить эти 10 ползунков всего двумя, назовем их "Общая Агрессивность" и "Стабильность Вождения". На выходе мы получили новую таблицу, где у каждой поездки вместо 10+ чисел было всего 2 (PC1, PC2). Теперь каждую поездку можно было отобразить как точку на обычном 2D-графике.

#### 3.3. Кластеризация методом K-Means: выявление поведенческих архетипов

* **Алгоритм:** K-Means был применен к данным после PCA.
* **Определение числа кластеров:** Оптимальное количество (4) было найдено с помощью "метода локтя".

##### Техническая реализация и Output Алгоритм нашел 4 центра скопления точек на нашем 2D-графике. Каждая поездка была отнесена к ближайшему центру. Мы проанализировали, какие исходные характеристики "ДНК" соответствуют каждому центру, и дали им бизнес-интерпретацию.Простыми словами об Output: В нашу таблицу с двумя колонками (PC1, PC2) добавилась третья — cluster\_id. Теперь у каждой поездки есть "бирка" с номером ее группы. Например, все поездки с cluster\_id = 2 — это "Агрессивные/Нервные водители", а с cluster\_id = 0 — "Спокойные городские водители". Мы автоматически разложили все поездки по полочкам.

#### 3.4. Обнаружение аномалий с помощью Isolation Forest

* **Цель:** Найти поездки, которые не похожи ни на один из типичных стилей.
* **Алгоритм:** Мы использовали Isolation Forest, так как он эффективно находит "белых ворон".

##### Техническая реализация и Output Алгоритм случайным образом "разрезает" пространство признаков, пока каждая точка не окажется в одиночестве. Аномалии, находящиеся на отшибе, изолируются гораздо быстрее.Простыми словами об Output: В нашу таблицу добавился еще один столбец — anomaly\_score. Это "индекс странности" поездки. У большинства поездок он низкий. Но если у какой-то поездки этот показатель высокий, это сигнал системе: "Обрати внимание, эта поездка была крайне необычной, она не похожа ни на один из стандартных стилей вождения".

#### 3.5. Результат Фазы 2: обогащенный поведенческий профиль поездки

Каждая поездка в нашем датасете была дополнена двумя новыми полями: cluster\_id (принадлежность к архетипу) и anomaly\_score. Это превратило исходный набор метрик в глубокий и многогранный поведенческий профиль.

### 4. Фаза 3: Гео-анализ и создание карты рисков

На этом этапе фокус анализа сместился с отдельных поездок на городскую среду. Цель — понять, **где** именно в городе проявляется рискованное поведение, и создать на основе этого карту рисков и спроса.

#### 4.1. Цель: переход от анализа "кто" к анализу "где"

Мы задались вопросом: рискованное вождение — это свойство только водителя или и городской среды тоже? Эта фаза отвечает на вторую часть вопроса.

#### 4.2. Геопространственная индексация H3: технология и обоснование выбора

* **Технология:** Мы использовали гексагональную систему геоиндексации H3. Представьте, что мы накрыли город медовой сотой. H3 — это система, которая дает каждой такой ячейке-соте уникальный адрес.
* **Выбор разрешения:** Мы выбрали **разрешение 9** (~175 метров), так как это оптимальный масштаб для городского анализа.

#### 4.3. Методология расчета risk\_score

risk\_score — это интегральный показатель, оценивающий, насколько "опасным" является вождение в данной H3-ячейке.

##### Техническая реализация и Output

* 1. **Агрегация на уровне ячейки:** Мы вернулись к датасету с точками (geotracks\_cleaned.parquet) и сгруппировали все точки по их H3-адресу. Для каждой ячейки рассчитали p95\_speed\_change, p95\_turn\_rate и stop\_density. Мы смотрим не на среднее поведение, а на самое экстремальное (p95), потому что именно оно приводит к инцидентам.
  2. Нормализация и взвешивание: Показатели были нормализованы и объединены по взвешенной формуле:  
     $ \text{risk\_score} = 0.4 \cdot \text{risk\_speed} + 0.4 \cdot \text{risk\_turn} + 0.2 \cdot \text{risk\_stop} $

**Простыми словами об Output:** На выходе мы получили таблицу, где каждая строка — это одна "сота" на карте города (например, h3\_id: 8928347...). А столбцы — это ее оценки: "уровень риска резких торможений", "уровень риска резких поворотов" и итоговый, единый risk\_score. Теперь у нас есть числовая оценка опасности для каждого маленького участка города.

#### 4.4. Методология расчета demand\_score

demand\_score отражает востребованность H3-ячейки как точки назначения.

##### Техническая реализация и Output

* 1. **Определение конечных точек:** Для каждой поездки была найдена ее последняя по порядку точка с помощью метода .tail(1). Это самый точный способ определить, где именно завершился маршрут.
  2. **Агрегация и нормализация:** Мы посчитали, сколько поездок закончилось в каждой H3-ячейке, и нормализовали это число.

**Простыми словами об Output:** Мы получили еще одну таблицу "адрес соты -> число". Например, h3\_id: 8928347..., demand\_score: 0.95. Это означает, что данный участок города является очень популярным местом назначения.

#### 4.5. Результат Фазы 3: геоповеденческий слой данных

Итогом фазы стал файл h3\_grid\_scores.geojson. Это мощный гео-слой. Мы объединили две наши таблицы и добавили к каждой "соте" ее географические координаты.

**Простыми словами об Output:** Финальный результат — это по сути цифровая карта. В ней для каждой "соты" (гексагона) на карте города записано два главных числа: насколько рискованно там водят (risk\_score) и насколько эта зона популярна как точка назначения (demand\_score). Этот файл полностью готов для визуализации.

### 5. Фаза 4: Прототипирование и "RL-Фасад"

Цель финального этапа — создать интерактивный инструмент, который наглядно демонстрирует результаты анализа и их бизнес-ценность.

#### 5.1. Цель: визуализация и демонстрация бизнес-применения

Мы создали веб-приложение (дашборд) для исследования профилей водителей и карты города.

#### 5.2. Технологический стек прототипа

* **Streamlit:** Для быстрого создания интерактивного веб-интерфейса.
* **Pydeck:** Для построения высокопроизводительной 3D-карты города.
* **Plotly:** Для создания интерактивных графиков.

#### 5.3. Компонент 1: Интерактивная карта рисков и спроса

##### Output для пользователяПользователь видит интерактивную 3D-карту Алматы. На ней город покрыт гексагонами. Красные и высокие "столбики" — это зоны, где водители ведут себя наиболее рискованно. Синие и высокие столбики — зоны высокого спроса. Можно переключаться между этими видами, приближать карту и, наводя мышку на любой гексагон, видеть точные цифры. Это позволяет менеджерам и аналитикам мгновенно находить самые проблемные или самые востребованные участки города.

#### 5.4. Компонент 2: Анализ "ДНК" водителя

##### Output для пользователяНа отдельной вкладке пользователь может ввести ID любой поездки. На экране появляется радар-чарт ("паутинка"). Эта "паутинка" наглядно показывает сильные и слабые стороны стиля вождения в данной поездке по сравнению со средними показателями. Например, можно сразу увидеть: "Ага, у этого водителя все показатели в норме, кроме экстремально резких поворотов".

#### 5.5. Инновация "RL-Фасад": имитация системы принятия решений

* **Концепция:** Мы создали функцию, которая имитирует работу "умного советчика" для водителей.

##### Техническая реализация и Output Функция берет любую ячейку, смотрит на 6 ее соседей и для каждого рассчитывает "оценку привлекательности" по формуле: $ \text{utility} = w\_d \cdot \text{demand\_score} - w\_r \cdot \text{risk\_score} - w\_c \cdot \text{distance} $Простыми словами об Output: В дашборде под картой для самой рискованной зоны автоматически выводится текстовая рекомендация. Например: "Для текущей ячейки с высоким риском (0.85) рекомендуется переместиться в соседнюю ячейку 892834... (на северо-восток), где риск значительно ниже (0.3), а спрос остается высоким (0.75)." Это демонстрация того, как система может в будущем давать водителям полезные подсказки в реальном времени.

#### 5.6. Результат Фазы 4: работающий интерактивный прототип

В результате был создан полнофункциональный дашборд, который не просто показывает данные, а позволяет с ними взаимодействовать, исследовать инсайты и демонстрирует конкретный пример их операционного применения.

### 6. Заключение и дальнейшее развитие

#### 6.1. Итоги и достигнутые результаты

Проект успешно прошел все запланированные фазы. Мы превратили сырые и зашумленные геотреки в многослойный аналитический продукт, включающий:

* Систему поведенческого профилирования поездок ("ДНК").
* Модель для сегментации стилей вождения.
* Уникальную геоповеденческую карту города с оценками риска и спроса.
* Интерактивный прототип, демонстрирующий бизнес-применение результатов.

#### 6.2. Бизнес-ценность и стратегическое применение

(Этот раздел полностью соответствует разделу 2 из краткого отчета, предоставленного ранее, и описывает ценность в области проактивной безопасности, снижения расходов, повышения эффективности и создания нового дата-актива).

#### 6.3. Направления для дальнейшего развития

Созданный движок является мощной платформой для дальнейших улучшений и внедрений:

1. **Улучшение ML-модели:** Переход от PCA к **Variational Autoencoders (VAE)**. Это позволит нам не просто группировать поездки, а создавать "генеративную" модель стилей вождения. Мы сможем, например, синтезировать пример "идеально безопасной" поездки по заданному маршруту или точнее оценивать отклонение от нормы.
2. **Внедрение полноценного RL-агента:** Замена "RL-Фасада" на реальную систему **обучения с подкреплением**. Такой агент будет не просто давать советы, а обучаться на реакции водителей. Если водители принимают совет и получают больше заказов, агент получает "награду" и понимает, что его стратегия верна, динамически адаптируясь к меняющейся обстановке в городе.
3. **Интеграция с продакшн-системами:** Встраивание risk\_score в алгоритмы ценообразования и распределения заказов. Разработка системы real-time алертов для службы безопасности.
4. **Расширение источников данных:** Обогащение модели данными о погоде, времени суток, типе дороги и дорожных событиях для еще более точного скоринга рисков.