Отчет по проекту:

Поиск фрода в СитиМобил

# Команда:

Дмитрий Черников

Олег Черемисин

Алексей Мышлянов

Константин Елизаров

Павел Шальнов

# Постановка задачи

Водители, работающие в Ситимобил, получают различные денежные выплаты, в том числе не связанные непосредственно с затратами на осуществление перевозки. Так, например выплаты могут начисляться за выполнение определенного числа поездок, или за поездки в определенных районах, а также в случае когда поездка слишком дешевая.

Чтобы увеличить свой доход недобросовестные водители пытаются обмануть сервис, используя нелегальные методы - **фрод**. Например, создают заказы сами себе, эмулируют поездки, отказываются от поездок в обход штрафных санкций и т.д.

Наиболее нетривиальным является задача выявление фрода, так где поездки физически вообще нет. Где и пользователь и водитель используют специальное программное обеспечение, чтобы эмулировать саму поездку.

Так называемый «фейк-GPS» российские таксисты начали использовать еще в 2016 году. Водители меняли координаты машины, чтобы оказаться ближе всего к клиенту. Подмена локации чаще всего использовалась для получения заказов из аэропортов, при этом реально таксисты находились в других районах. Однако, сейчас эмулируется вообще все, включая показания акселерометра, данные о местоположении с вышек сотовой связи и так далее.

Наша задача, - построить модели, позволяющие размечать фейковые поездки и водителей.

Рынок такси в России по итогам 2021 года аналитики оценивают в 866 млрд рублей, ожидая рост на 6% в годовом выражении. К 2025 году этот показатель достигнет 1,06 трлн рублей, считают эксперты. Большая часть такси в России заказывается через агрегаторы, на них приходится 60% от всех заказов. Убытки агрегаторов связанные с мошенничеством составляют миллиарды, если найти даже 5% это очень много денег.

Поиск фрода является актуальной задачей не только в службах такси. Возможные финансовые убытки от мошенничества несут также набирающие популярность службы доставки, курьерские службы.

Анализ геоданных и выявление аномалий представляет интерес для многих направлений, связанных с организацией перевозок и перемещением товаров и грузов.

# Обзор известных методов

В настоящее время в моделях выявления фрода в такси в основном используются данные о регистрации водителей и пассажиров, идентификаторы смартфона и SIM, данные спутникового и мобильного геопозиционирования, платежные данные пользователей.

Наиболее частым является Rule-based подход, который заключается в составлении совокупности правил для выявления отклонений в данных, соответствующих шаблонам мошеннических действий. Rule-based подход используется в России как крупными агрегаторами, такими как Ситимобил, Яндекс Такси, так и мелкими, а также, например, крупным американским Lyft.

Известно, что мировой лидер по такси-перевозкам, компания Uber для выявления фрода использует данные сообщества миллионов пользователей, как водителей, которых насчитывается более 5 миллионов, так и пассажиров, которых порядка 100 миллионов. Система антифрода в Uber построена с применением графовой сверточной сети (Relational Graph Convolutional Networks).

Сведения по антифроду в крупных агрегаторах Европы (Bolt, Gett), Китая (Didi), Индии (Ola Cabs) и стран Юго-Восточной Азии (Grab, Go-Jek) в свободном доступе отсутствуют. Но опубликованные на их сайтах вакансии говорят о том, что эти компании активно набирают свои ML-команды и развивают DS-направление для поиска андтифрода.

Задачи машинного обучения в антифроде в такси две: первая - классификация и кластеризация водителей и пассажиров на основе категориальных и числовых признаков; вторая - поиск аномалий во временных рядах с историями поездок.

Наш проект нацелен на решение второй задачи, это задача **Trajectory Outlier Detection.** Различные научные подходы решение данной задачи регулярно публикуются с 2008 года. В рамках работы над проектом подготовлен обзор методов поиска аномалий в геоданных.

Основные используемые подходы:

* Distance-based алгоритмы используют вычисление расстояний между соседними точками, такие алгоритмы эффективны для определения похожести траекторий.
* Density-based алгоритмы используют методы кластеризации на основе пространственной плотности (DBSCAN).
* Pattern mining-based алгоритмы преобразуют массив траекторий в массив транзакций для адаптации к различным алгоритмам поиска шаблонов (pattern mining).
* Алгоритмы, использующие методы машинного обучения, такие как SVM, PCA, уменьшение размерности, а также различные их ансамбли.
* Алгоритмы скоринга, которые выделяют выбросы сравнивая для траекторий различные метрики.

# Описание проекта

Цели проекта:

* исследование существующей литературы по Fake-GPS и поисках аномалий в данных геолокации, а также в любых других связанных данных;
* построить ML-модель, которая на основе логов поездок, содержащих показания датчиков GPS и акселерометра, находит подозрительные поездки, а также делает скоринг водителей на вероятность фрода.

Инструменты:

В requirements традиционный ML, Deep Learning на Pytorch, из специфики - Folium, OSMPythonTools. Код в Jupyter ноутбуках. Демонстрация выполнена на платформе Streamit. Репозиторий проекта выложен на GitHub.

Данные по каждому водителю представлены в виде двух датасетов - логи трекера МайТрекер (координаты, время) и логи приложения Ситимобил (координаты, время и показания акселерометра).

Данные абсолютно реальные, там есть технические баги, сбои устройств, пропущенные значения в логах и тп.

Имеющиеся данные дают большое пространство для фич-инжениринга. Здесь можно выделить следующие категории признаков.

**Признаки местоположения** - зоны / районы / кластера, удаленность от города / центра / объектов городской инфраструктуры, наложение на дорожную сеть / карту городской застройки.

**Временный признаки** - интервалы в движении / точки остановок, периодичность в течение дня / недели.

**Линейные величины** - расстояние / скорость / ускорение, их статистики, агрегации и прочие производные.

**Угловые величины** - углы поворота / угловая скорость, рысканье (тангаж / крен)

Линейные и угловые величины можно измерять между последовательными точками, поточечно с некоторым шагом, усреднять окном. Также имеет смысл делать агрегацию точек в рамках каждой поездки или за определенный интервал времени.

Следует отметить, что использование признаков местоположения и вообще самих координат ограничивает применение модели в пределах одной локации (город, район), и для переноса модели на другую локацию потребуется переобучение.

# Описание моделей

Поскольку данные представляют собой два датасета, полученные из разных источников - логи трекера и логи антифрода, мы исследуем возможность детекции аномалий по каждому отдельному датасету, результаты полученных моделей вместе с совокупностью данных будут использованы для окончательной классификации поездок на фрод или не фрод, а также для возможного скоринга водителей. Сложность тут в том, чтобы отличать аномалии связанные с фродом и аномалии, связанные с работой технических устройств (например, вышек связи, глушилок и тому подобное).

Другую сложность составляет то, что как и в общем случае при решении задачи детекции аномалий мы имеем дело с обучение без учителя, где мы не можем заранее сформулировать признаки аномалий. В данном направлении мы также проводим моделирование и эксперимент.

**Random Forest / Catboost**

**LSTM**

Для определения фрода на уровне каждой поездки была построена рекуррентная нейросеть LSTM. Архитектура сети представлена на схеме внизу слайда.

На вход каждого блока подаются конкатенированные данные GPS и акселерометра, с рассчитанными скоростями и прочими признаками вроде приращения координат. Итоговая модель обладает достаточно высокой точностью предсказания, порядка 82 процентов.

В этом можно убедиться, взглянув на рисунок справа. Модель хорошо отделяет фродовые поездки от нефродовых, притом детектит достаточно малую часть неразмеченных поездок со "странным" в понимании модели поведением.

Итоговые метрики следующие IoU 0.028, что значительно ниже, чем у остальных моделей. Это объясняется тем, что нужно подкорректировать пороги отсечения по предсказанным вероятностям. Корреляция равна 0.12, что является хорошим результатам, и означает, что модель способна адекватно ранжировать поездки, а следовательно и водителей, по факту фрода.

**Isolation Forest**

Изолирующий лес, логика алгоритма простая: при описанном «случайном» способе построения деревьев аномалии будут попадать в листья на ранних этапах (на небольшой глубине дерева).

Перемен

Пробовали отлавливать аномалии на поездках и на точках, по поездкам поиск аномалий не дал нормального результата. На точках же, процент аномальных точек для фродовских водителей составил 100 %. Большинство точек точно НЕ фродовских водителей, лежит в пределе 30 %. Выбрали порог в 89 %, для определения фрода, так как последний НЕ фродовский водитель имеет порог в 85 %. Скор балл получился соответственно из процента аномальных точек по водителю.

**k-Means**

В качестве модели unsupervised подхода реализована k-means модель. Модель обучалась на данных GPS с группировкой точек в поездки, паузы между интервалами времени которых меньше 5 минут. Выполнялась предобработка данных и генерация фичей, как агрегатов по поездкам (например продолжительность поездки, длина, максимальные, медианные, кумулятивные значения скорости и ускорения), так и общие фичи (будний/выходной день, флаг часа пик, кодировка времени тригонометрией).

Подбор оптимального количества кластеров осуществлялся по тому, насколько однородные получаются кластера по соотношению долей фродовых и не фродовых точек. Лучший результат показала модель с 5 кластерами. На рисунках можно увидеть, что кластера под номерами 1 и 3 достаточно хорошо отделились от общей массы точек и не содержат в себе точек не фродовых поездок, что позволяет классифицировать неизвестные точки, попавшие в данные кластера, как фродовые. Также выделяется кластер 0 с полным отсутствием фродовых точек, однако туда и не попали точки подлежащие классификации. В итоговом предсказании 2 водителя были классифицированы как фродовые в результате попадания их точек в кластеры 1 и 3. Лишь небольшая часть точек попала в данные однородные кластера, основная масса точек лежит в оставшихся двух самых больших кластерах.

# Результаты

Модели показали разный результат. Большая часть предсказаний отличается от результатов полученных командой Ситимобил. Считаем, что причина в отсутствии GroundTruth, на текущий момент не разработаны еще модели, реально описывающие перемещение водителя и позволяющие выделять фрод.

Демонстрация предсказаний разработанных проектов выполнена на платформе Streamlit:

<https://share.streamlit.io/cherninkiy/made21_fraud_detection_project/main/demo/demo.py>

В ходе работы над проектом:

* подготовлен обзор существующей литературы по Fake-GPS и поисках аномалий в данных геолокации;
* проведен Exploratory Data Analisys, выявлены пропуски и недочеты, реализованы методы их устранения
* по исходным данным сформированы признаки из разных, описанных ранее, категорий, изучены их статистические свойства
* реализованы четыре модели на различных архитектурах, рассмотрены подходы supervised / unsupervised learning
* проведены эксперименты, получены результаты, подготовлена демонстрация

Репозиторий проекта размещен на GitHub:

<https://github.com/cherninkiy/made21_fraud_detection_project>