В. О. Сандуян, А. В. Семененко, Т. Ю. Чернышева, Л. Н. Бакановская
Тюменский государственный университет, г. Тюмень
УДК 004.85

АНАЛИЗ ДАННЫХ СЕРВИСА ТАКСИ С ПРИМЕНЕНИЕМ МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Аннотация. В рамках данной статьи проведен анализ данных поездок такси с использованием моделей искусственного интеллекта [5] для выявления оптимальных стратегий по снижению издержек и увеличению прибыли для компании. Результаты исследования представлены в виде конкретных рекомендаций заказчику, направленных на повышение эффективности и качества обслуживания.

Ключевые слова: такси, анализ данных, искусственный интеллект, оптимизация, гипотезы, k-means.

Введение. С развитием цифровых технологий в сфере транспорта данные о поездках такси становятся все более доступными и широко используемыми для анализа. Использование данных поездок такси открывает возможности для выявления основных проблем, оптимизации маршрутов и повышения эффективности обслуживания клиентов [6]. В рамках исследования ставится задача анализа данных о поездках с применением моделей искусственного интеллекта [2], с целью выработки гипотез по оптимизации процессов и повышения эффективности обслуживания клиентов заказчика — компании такси [4].

Проблема исследования. Решаемая проблема — неэффективное использование ресурсов и неоптимальные маршруты в работе такси, приводящие к излишним издержкам и недостаточной прибыли, требующие применения инновационных подходов на основе анализа данных и использования методов машинного обучения.

Целью работы является анализ данных о поездках такси с применением моделей искусственного интеллекта для выявления неэффективных практик в планировании маршрутов и использовании ресурсов компании, а также формирование рекомендаций для снижения издержек, увеличения прибыли и повышения качества обслуживания клиентов.

Материалы и методы. Для первоначальной обработки и анализа данных о поездках такси использована библиотека Pandas на Python. Представленный заказчиком датасет состоит из 9 столбцов:

- **id —** идентификатор замера геоданных;
- **id_order —** идентификатор заказа;
- **id_driver —** идентификатор водителя;
- **c_date —** дата и время замера геоданных;
- **c_lat —** местонахождение водителя в момент замера геоданных (широта);
- **c_lon —** местонахождение водителя в момент замера геоданных (долгота);
- **c_vector —** вектор направления водителя;
- **c_precision —** точность замера геоданных.

На основе этих данных сформулирована основная часть гипотез.

В результате анализа установлено, что датасет относится к временному периоду с 28.12.2023 по 11.01.2024 и содержит 10 242 100 строк, среди которых записи о 223 154 заказах и 4490 водителях. Таким образом, замеры геоданных производились в среднем 45 раз за каждую поездку. При этом в среднем на каждого водителя приходится около 49 заказов, но медианное значение равно 26. Разница между медианным и средним значениями свидетельствует

о наличии выбросов, в данном случае речь идет о водителях, выполнивших наибольшее и наименьшее число заказов (рис. 1).

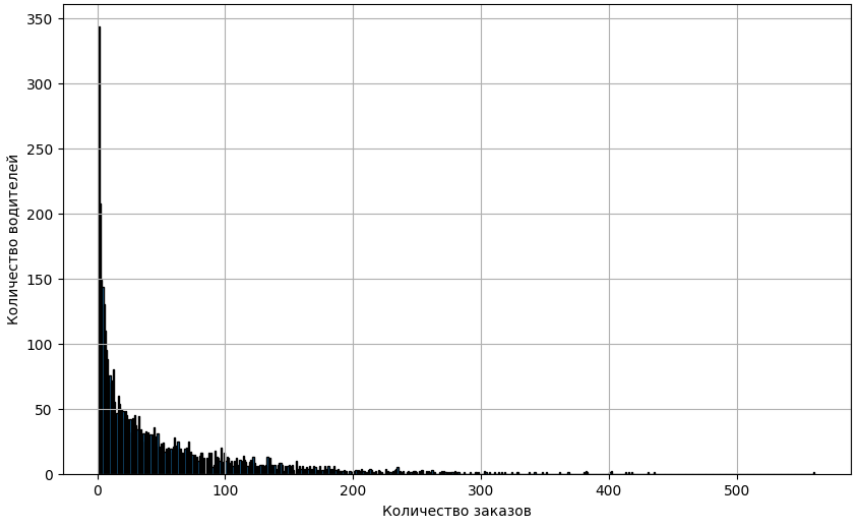

Рис. 1. Распределение нагрузки между водителями

Дальнейший анализ работы водителей показал, что 552 водителя (12% от общего числа) выполнили менее 3-х заказов за две недели. В то же время 5 водителей, выполнивших наибольшее число заказов, за тот же период осуществили 2261 поездку (в среднем 452 заказа на одного водителя). Объяснение разницы между данными водителями стало дальнейшим шагом, способствующим формированию рекомендаций.

Для данной задачи использовались инструменты визуализации Python (библиотека Folium), а также геоинформационная система QGIS [3]. Выявлено, что географическое местоположение данных относится к поездкам внутри города Курган, а также в соседние города (рис. 2).

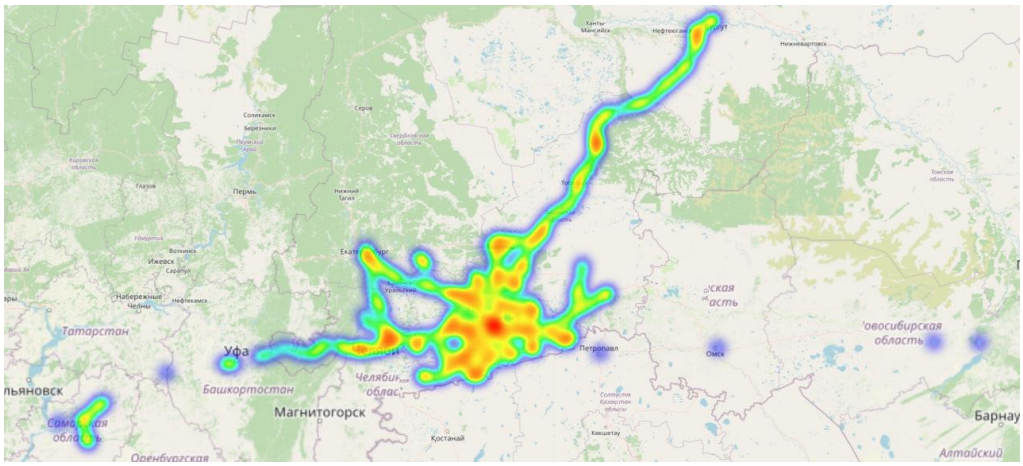

Рис. 2. Географическое местоположение данных

При анализе передвижения 5 водителей, выполнивших наибольшее число заказов, установлено, что все они осуществляли поездки в пределах города, в то время как водители, выполнившие наименьшее число заказов, осуществляли поездки между городами (см. рис. 3).

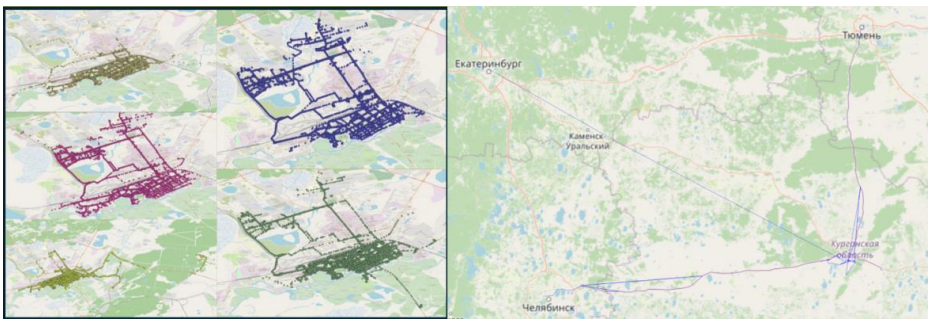

Рис. 3. Передвижения водителей, выполнивших наибольшее и наименьшее число заказов

Более того, авторами применен алгоритм кластеризации k-means, который является популярным методом машинного обучения для группировки данных.

Число кластеров выбрано на основе метода локтя в количестве 26 (рис. 4).

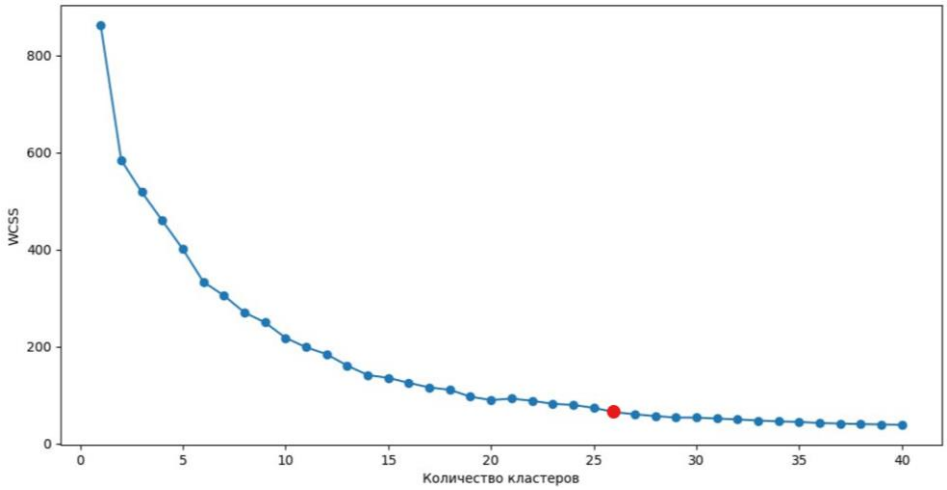

Рис. 4. Метод локтя для определения количества кластеров

Кластеризация позволила выявить зоны высокого спроса в определенное время суток. Результаты кластеризации проверены методом силуэтов [7] (рис. 5).

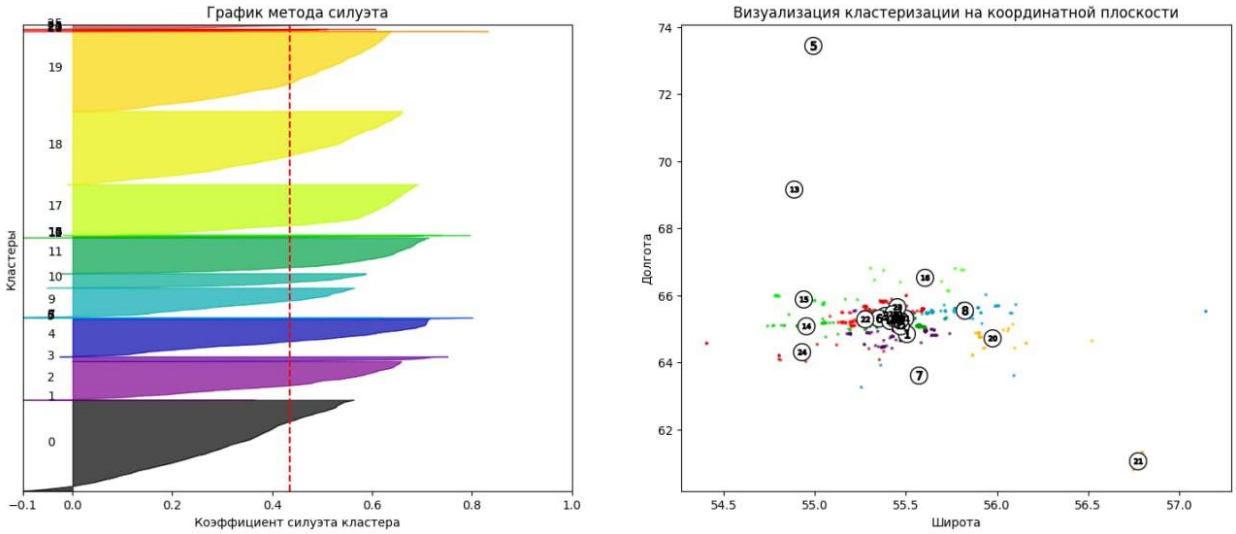

Рис. 5. Метод локтя для определения количества кластеров

Таким образом, на основании предоставленного графика можно сделать вывод, что среднее значение коэффициента силуэта для оцениваемого алгоритма кластеризации составляет около 0,42, что указывает на умеренную степень соответствия точек данных их назначенным кластерам. Однако важно также рассмотреть распределение коэффициентов силуэта. Коэффициенты силуэта для оцениваемого алгоритма кластеризации достаточно высокие, причем большинство коэффициентов составляет от 0,5 до 1. Это говорит о том, что алгоритм кластеризации хорошо выполняет назначение точек данных соответствующим кластерам.

Результаты. Анализ данных о поездках такси позволил сформировать следующие гипотезы:

1) «Центр притяжения» — зависимость между вектором движения, широтой, долготой и популярными местами в городе. Водители чаще могут двигаться в сторону аэропорта, ТЦ или другого места с пиковым спросом на такси. Анализ данной зависимости поможет определить места пикового спроса в городе.

2) «Зона высокого спроса в определенное время» — связь между широтой, долготой и временем описывает изменение координат водителя в определенный час/день. Анализ зависимости поможет определить, какие места являются популярными среди клиентов в определенное время суток.

3) «Объезд пробок» — построив маршрут водителя, который проезжает через «центр притяжения» либо зону высокого спроса, можно сравнить вектором движения и скорость, тем самым определив, с какой стороны быстрее подъехать к клиенту.

4) «Характер водителя» — определенные водители имеют собственные предпочтения по скорости движения и направлению. Можно проверить, существует ли разница в скорости движения между разными водителями. Это влияет на безопасность, время и маршрут выполнения заказов.

Визуализация маршрутов поездок и точек заказов позволила наглядно представить данные о поездках и выделить основные тренды и аномалии. Авторами определены «центры притяжения». В разное время дня ими стали:

1) Утро: Аэропорт, Отели/Гостиницы, Больницы/Поликлиники.
2) День: Отели/Гостиницы, Вокзал, ТЦ, Рестораны/Столовые.
3) Вечер: Рестораны/Столовые, Вокзал.
4) Ночь: Аэропорт.

Проанализированы данные по числу водителей в определенный час каждой даты: лучшие (рис. 6) и худшие часы по количеству заказов (рис. 7).

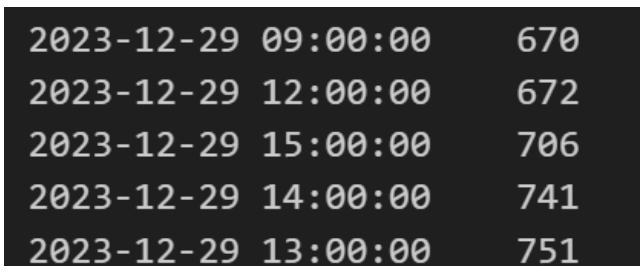

Рис. 6. Лучшие часы по количеству заказов

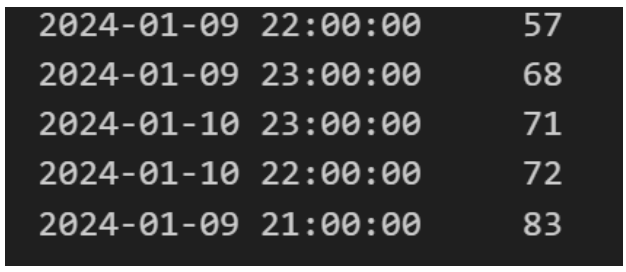

Рис. 7. Худшие часы по количеству заказов

Самым неактивным периодом времени были часы с 21:00 до 03:00, поэтому необходимо поработать над улучшением алгоритма подбора цены заказа в ночное время. Выяснилось, что водители, выполнившие наибольшее число заказов, работают сменами 12 часов каждый день с 03:00 до 15:00. А при анализе работы водителей, выполнивших всего 1 заказ, установлено, что почти все они работают на междугородние поездки. Без данных о стоимости поездки сложно сказать, почему они не хотят работать в городе с заказчиком. Возможно, комиссия на выполнение одного междугороднего заказа значительно меньше именно у заказчика, а городские заказы выгоднее в других агрегаторах, поэтому рекомендуется поработать над retention-маркетингом (удержание и вовлеченность) таких водителей.

Применение алгоритма кластеризации k-means на основе географических данных позволило определить зоны высокого спроса в определенное время (рис. 8), а также центры выявленных кластеров (рис. 9).

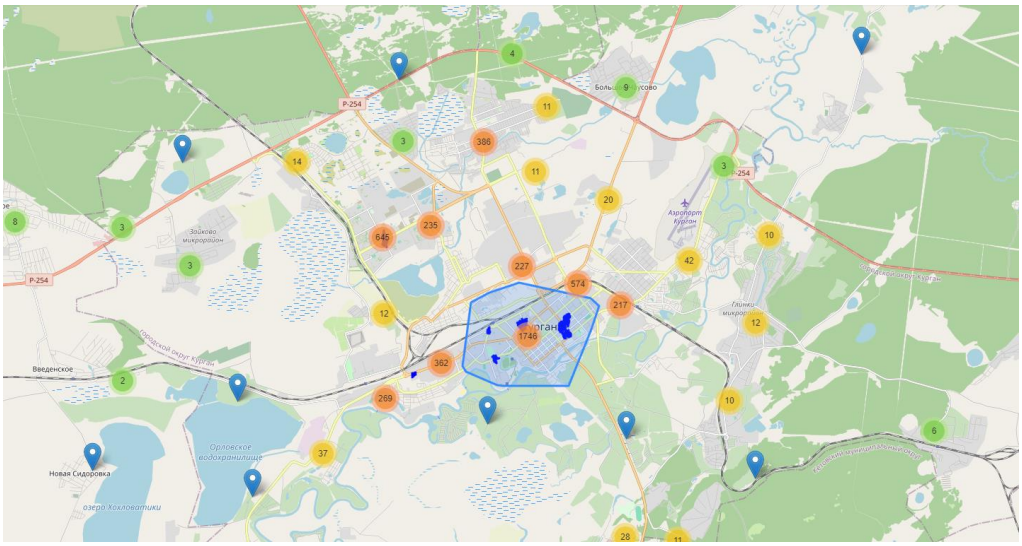

Рис. 8. Визуализация зон высокого спроса на карте города

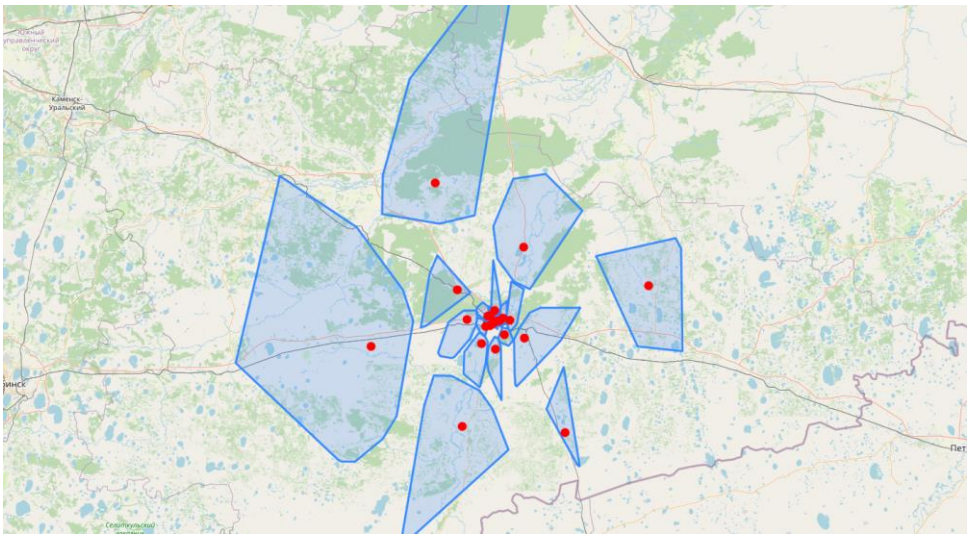

Рис. 9. Центры кластеров, являющихся зонами высокого спроса, на карте

Заключение. В ходе исследования проведены анализ данных о поездках такси, визуализация маршрутов и применение алгоритма кластеризации k-means на основе географических данных. Полученные результаты позволили выявить неэффективные практики в планировании маршрутов и использовании ресурсов компании, а также выделить оптимальные стратегии для улучшения работы компании.

Стояла задача предоставить заказчику конкретные рекомендации по оптимизации процессов. На основе проведенного анализа и кластеризации рекомендуется:

1) оптимизировать алгоритм определения цены заказа в ночное время (с 21:00 до 03:00), поскольку количество водителей существенно сокращается;

2) поработать над retention-маркетингом водителей, работающих на междугородние заказы;

3) внедрить алгоритм определения зон высокого спроса в определенное время (например, методом кластеризации k-means) для предсказания заказов;

4) использовать центроиды кластеров как места ожидания заказов. Это оптимальные места, паркуясь на которых, можно быть равноудаленным от границ определенного кластера;

5) использовать центроиды кластеров для оптимизации маршрутов. Рассчитывая регулярные маршруты вокруг центроидов кластеров, водители такси могут эффективно обслуживать клиентов, уменьшая время в пути и затраты на топливо;

6) анализировать динамику изменения кластеров во времени. Проводить периодический мониторинг данных и обновление модели кластеризации, чтобы адаптировать стратегию работы такси под изменяющиеся условия и потребности клиентов;

7) использовать кластеризацию для определения мест с разным уровнем спроса и оптимального распределения количества такси по районам;

8) рассмотреть возможность внедрения акций для повышения привлекательности такси в разных кластерах. Например, предложение скидок в малопосещаемых районах или разработка специальных тарифов для предпочитающих определенные маршруты клиентов.

Таким образом, реализация предложенных рекомендаций поможет компании улучшить свою деятельность, повысить конкурентоспособность и удовлетворить потребности клиентов.

СПИСОК ЛИТЕРАТУРЫ

1. Применение интеллектуального анализа данных для прогнозирования / Д.А. Ворсина, С.А. Нестеров. — Текст: электронный // Системный анализ в проектировании и управлении. — 2019. — № 1. — URL: https://cyberleninka.ru/article/n/primenenie-intellektualnogo-analiza-dannyh-dlya-prognozirovaniya (дата обращения: 06.05.2024).
2. Сравнение алгоритмов кластеризации / Е.В. Лихтина. — Текст: электронный // Актуальные проблемы авиации и космонавтики. — 2018. — № 14. — URL: https://cyberleninka.ru/article/n/sravnenie-algoritmov-klasterizatsii (дата обращения: 06.05.2024).
3. Анализ дорожно-транспортных происшествий в городе Саранске с использованием QGIS / С.С. Беднов, О.Ф. Богдашкина, Л.Г. Калашникова. — Текст: электронный // Огарёв-Online. — 2023. — № 2 (187). — URL: https://cyberleninka.ru/article/n/analiz-dorozhno-transportnyh-proisshestviy-v-gorode-saranske-s-ispolzovaniem-qgis (дата обращения: 06.05.2024).

4. Методы определения требуемого количества легковых автомобилей такси на примере городов Республики Башкортостан / М.В. Дашко, В.П. Славненко, Е.Ю. Кириллов. — Текст: электронный // Вестник ОГУ. — 2015. — № 1 (176) — URL: https://cyberleninka.ru/article/n/metody-opredeleniya-trebuemogo-kolichestva-legkovyh-avtomobiley-taksi-na-primere-gorodov-respubliki-bashkortostan (дата обращения: 06.05.2024).
5. Алгоритмы машинного обучения / Д.А. Макаров, А.Д. Шибанова. — Текст: электронный // Теория и практика современной науки. — 2018. — № 6 (36). — URL: https://cyberleninka.ru/article/n/algoritmy-mashinnogo-obucheniya (дата обращения: 06.05.2024).
6. Алгоритм нахождения оптимального расположения распределительных центров для доставки заказов с использованием методов неконтролируемого машинного обучения / А.С. Акулов, А.Д. Безотосная, Т.В. Пак. — Текст: электронный // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. — 2023. — № 3. — URL: https://cyberleninka.ru/article/n/algoritm-nahozhdeniya-optimalnogo-raspolozheniya-raspredelitelnyh-tsentrov-dlya-dostavki-zakazov-s-ispolzovaniem-metodov (дата обращения: 06.05.2024).
7. Упрощенный показатель силуэта для определения качества кластерных структур / В. В. Журавлева, А. С. Маничева. — Текст: электронный // Известия АлтГУ. — 2022. — № 4 (126). — URL: https://cyberleninka.ru/article/n/uproschennyy-pokazatel-silueta-dlya-opredeleniya-kachestva-klasternyh-struktur (дата обращения: 06.05.2024).