DAN

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

ТИНЬКОФФ

ЗДРАВСТВУЙТЕ, УВАЖАЕМЫЕ УЧАСТНИКИ!

Сегодня мы предлагаем вам поближе познакомиться с удивительным миром анализа данных и поработать над проектом, приближенным к реальному, в команде с такими же исследователями (да, решение задач по анализу данных — командная работа!).

Исследование — это сложный процесс, поэтому для того, чтобы вы могли справиться со всеми возникающими трудностями и вопросами, к каждой команде будет прикреплен ментор.

Кто такой ментор?

Кто такой ментор? Это человек, который имеет опыт работы над исследовательским проектом, не раз сталкивался со всеми трудностями при решении аналитических задач, умеет их преодолевать;
а теперь готов помочь вам сделать качественный проект и победить!
Не стесняйтесь обсуждать с ментором любые свои идеи, предлагать ему различные способы решения поставленной задачи и спрашивать совета в ходе работы. Каждый ментор много знает и умеет, не стоит игнорировать его помощь, однако не надейтесь, что он сделает работу за вас! :)

ВВЕДЕНИЕ

Просмотр фильмов и сериалов дома с использованием различный приложений становится всё более популярной заменой традиционному походу в кино или просмотру телевизионных программ. И многие люди при выборе того, что смотреть часто ориентируются на рекомендации самого приложения. В исследовании вам будет предложена база взаимодействия пользователей с контентом приложения МТС Kion. За 2021-2022 г. в России пользователи совершили более 6 млн. взаимодействий (просмотр контента пользователем) из которых было случайно отобрано 98,367 значений.

Описание переменных

Доступная вам база данных включает следующие переменные
(в данных могут быть пропуски):

user_id — ID пользователя

age — возрастная группа пользователя, строка вида «M_N».

age_18_24 — от 18 до 24 лет включительно

age_25_34 — от 25 до 34 лет включительно

age_35_44 — от 35 до 44 лет включительно

age_45_54 — от 45 до 54 лет включительно

age_55_64 — от 55 до 64 лет включительно

age_65_inf — от 65 и старше

sex — пол пользователя

М — мужчина

Ж — женщина

income — среднемесячный доход пользователя в тыс.рублей, строка вида «M_N» (интервал для дохода)

income_0_20

income_20_40

income_40_60

income_60_90

income_90_150

income_150_inf

kids_flg — флаг «наличие ребенка» у пользователя (1, если есть; 0 — иначе)

item_id — ID контента

content_type — тип контента (фильм или сериал)

title — название на русском

genres — основной жанр фильма или сериала

countries — страна или страны (перечислены через запятую), где фильм или сериал был произведен

release_year — год производства фильма

for_kids — флаг «контент для детей» (1 — да, 0 — нет; есть пропущенные значения)

age_rating — возрастной рейтинг (0 — для всех возрастов; 6 — 6+ лет; 12, 16, 18, 21 — по аналогии)

watched_pct — сколько процентов фильма или сериала просмотрено

last_watch_dt — Дата последнего просмотра

total_dur — Общая продолжительность всех просмотров данного фильма/сериала в секундах

ЗАДАНИЕ

Вам нужно подготовить исследование взаимодействия пользователей с платформой. Результатом исследования должны быть дальнейшие рекомендации пользователям и/или кинотеатрам относительно фильмов/сериалов, основанные на предыдущих действиях пользователей (истории их взаимодействия с платформой). Исследование должно содержать:

Исследование должно содержать:

- Постановку вопроса и гипотезы. Должно быть высказано предположение или идея, которое будет проверяться в рамках исследования. Также необходимо обоснование логичности предложенной гипотезы (через предварительный разведывательный анализ или цепочку причинно-следственной связи A-> B->C->....);

  Анализ ситуации. Нужно дать понять зрителям, с какими данными вы работали, какие там есть особенности и с чем они связаны;

- Проверку гипотезы. При помощи статистического анализа нужно сделать вывод о подтверждении гипотезы или её опровержении в результате исследования (отрицательный результат тоже может приводить к выводам и рекомендациям);

- Вывод и рекомендации. Исследование должно быть ценно не само по себе, а вследствие возможностей его применения и полученных результатов;

- Перспективы развития. Результаты исследования должны быть критически оценены, рассмотрены ограничения, варианты улучшения в дальнейшем, и прокомментирована обобщаемость полученных выводов.

КОММЕНТАРИИ К ЗАДАНИЮ

1. **Постановка вопроса. Подумайте о том, что вам интересно исследовать и почему вам это интересно.** Представленный набор данных позволяет исследовать различные аспекты действий пользователя, его выбор, его предпочтения. На первом этапе обсудите вопрос, на который вы хотите получить ответ с использованием данных. При постановке вопроса подумайте, какая переменная (будем называть ее объясняемой переменной) отражает ту особенность, которую вам интересно изучить. Объясняемых переменных может быть несколько (тогда надо понять преимущества и недостатки использования каждой из них), или это может быть композитная переменная, составленная из исходных переменных.

   a. При анализе можно сконцентрироваться на исследовании фильмов или пользователей.

   b. Если вы изучаете фильмы, то при исследовании просмотров нас могут интересовать их продолжительность, очередность, а также их частота.

   c. Если вы изучаете пользователей, то будет интересно посмотреть на наборы фильмов, которые просматривают пользователи со схожими характеристиками (для этого нужно провести предварительные расчеты дополнительных переменных).

   При постановке вопроса подумайте, какая переменная (будем называть ее объясняемой переменной) отражает ту особенность, которую вам интересно изучить.

2. **Выбор переменных.** От чего может зависеть переменная, которую вы выбрали в предыдущем пункте? Выберите из представленного набора данных факторы, влияние которых вы считаете важным. Сформулируйте гипотезу о том, положительно или отрицательно эти факторы будут влиять на объясняемую переменную. Расскажите подробно, почему ожидаете именно такую зависимость. Подумайте, достаточно ли данных для объяснения гипотезы и, если нет, обдумайте, сможете ли вы найти и проанализировать недостающие данные (да, другие источники данных тоже можно использовать). Подумайте — как эти внешние источники соотносятся с представленной базой данных, а также какую информацию из них можно было бы добавить.

   При работе с данными обязательно обратите внимание на то, что является единицей наблюдения, т. к. это будет определяющим для всего вашего анализа.

   В представленном наборе данных единица наблюдения определяется взаимодействием пользователя с онлайн-кинотеатром. Это дает возможность рассматривать в качестве единицы наблюдения и само взаимодействие пользователя, и посетителя кинотеатра, и фильм/сериал.

3. **Разведывательный анализ.** Проиллюстрируйте с помощью графиков распределения двух переменных (зависимой и объясняющей), на которых вы решили сфокусироваться, их особенности. Посчитайте основные описательные статистики для них. Как можно проинтерпретировать, т. е. объяснить, наблюдаемые особенности и полученную статистику?

4. **Анализ взаимосвязей.** Посмотрев на переменные отдельно, можно переходить к проверке гипотезы. Отразите графически взаимосвязь между выбранными переменными (между зависимой и объясняющей, но иногда важно проследить и взаимосвязь между несколькими объясняющими) и проведите статистический анализ (обратите внимание на типы переменных, для разных типов инструменты различаются). Согласуется ли результат с поставленной гипотезой? Если нет, объясните, с чем это может быть связано или попробуйте посмотреть другие переменные, возможно они лучше объяснят вашу гипотезу. Проверьте, являются ли результаты устойчивыми при анализе по подгруппам? Например, получаются ли похожие результаты по фильмам с различным возрастным цензом? Проанализируйте исследуемую взаимосвязь, поделив выборку по какому-либо фактору. Объясните важность этого фактора. Как можно проинтерпретировать полученные результаты?

5. **Выводы и ограничения.** Какой интерес ваш результат представляет для данного онлайн-кинотеатра? Для других похожих платформ? Для кого еще он может оказаться полезен и чем? Подумайте и объясните, как пользователи, организации или органы государственной власти могут использовать полученные вами результаты.

   Идеальных исследований не бывает! Какие важные факторы, влияющие на вашу объясняемую переменную не были вами учтены (из представленного набора или из тех, что могут содержаться в других источниках или среди факторов, которые сложно объективно измерить)? Насколько трудоёмко получить такие дополнительные данные? Что можно сказать о самой предложенной, о ее репрезентативности? Можно ли распространять полученные выводы на все онлайн-кинотеатры (или пользователей) России, других стран и почему? Наконец, обсудите, возможна ли обратная причинно-следственная связь в вашей гипотезе. Если да, объясните, почему, а если нет — придумайте экономический обоснование для обратного влияния.