DAN

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

ТИНЬКОФФ

ЗДРАВСТВУЙТЕ, УВАЖАЕМЫЕ УЧАСТНИКИ!

Сегодня мы предлагаем вам поближе познакомиться с миром анализа данных и поработать над реальной задачей в команде с такими же юными исследователями.

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, 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 всю работу за вас! :)

ВВЕДЕНИЕ

Посещение кафе и ресторанов стало неотъемлемой частью жизни жителей России. Для удобства выбора существует система рейтингов и оценок, которые пользователи могут оставлять посещенным местам. При выборе кафе или ресторана люди часто ориентируются на рейтинги, составленные предыдущими посетителями, что делает рейтинги важными и для самих организаций. В исследовании вы рассмотрите отзывы, оставленные жителями Москвы и Санкт-Петербурга на предприятия общественного питания (далее — организации), работающие в соответствующих городах. Так, за 3 года, начиная с февраля 2018, в обоих городах жителями было оставлено свыше 3 млн. отзывов, из которых было случайно отобрано 104,278. Заметим, что пользователи могут как посещать организации в своем городе (жители Москвы — в Москве, жители Санкт-Петербурга — в Санкт-Петербурге), так и путешествовать и оставлять отзывы в чужом (москвичи в организациях Санкт-Петербурга и наоборот).

Данные об отзывах (оценка, город и дата) были объединены с данными об организациях (средний рейтинг, средний чек, город и направления/особенности деятельности).

Описание переменных

Доступная вам база данных включает следующие переменные:

1. user_id — идентификатор пользователя (93,786 уникальных значений).

2. org_id — идентификатор организации (22,925 уникальных значений).

3. rating — поставленная оценка в отзыве (от 1 до 5).

4. ts — дата отзыва (последовательный номер дня, начиная с 01.02.2018 — это 0 день).

5. user_city — информация о городе проживания пользователя (msk — Москва, spb — Санкт-Петербург).

6. org_city — информация о городе организации (msk — Москва, spb — Санкт-Петербург).

7. average_bill — средний чек в рублях (округленный с точностью до 500 рублей).

8. rating_org — средний рейтинг организации (от 1 до 5).

9. rubrics — рубрика, указывающая на основной вид деятельности организации (например, «Ресторан», «Кафе», «Быстрое питание» и т. п.).

10. food_delivery:handmade_goods — 63 последних столбца в таблице представляют собой дамми-переменные (или бинарные переменные), связанные с особенностями организаций. Каждая из них принимает значение 1, если данная особенность относится к организации, и 0, если нет. Например, если переменная breakfast равна 1, это означает, что в организации подают завтраки. Как правило, множество особенностей организации определяется ее владельцем и может быть неточным.

Данные получены с «Яндекс Карт», так что вы можете посмотреть, как они выглядят для пользователя в самом сервисе, выбрав какую либо из организаций на карте.

ЗАДАНИЕ

Вам нужно подготовить исследование оценок, поставленных организациям. Результатом исследования должны быть рекомендации по дальнейшей политике: бизнесу (политика поведения) и/или регулятору, т. е. государству (политика регулирования).

Исследование должно содержать:

- Постановку вопроса и гипотезы. Должно быть высказано предположение, которое будет проверяться в рамках исследования, и дано обоснование адекватности выдвинутой гипотезы — это достигается через механизм гипотезы
- Анализ ситуации. Нужно дать понять слушателям, с какими данными вы работали, какие там есть особенности и с чем они связаны
- Проверку гипотезы. При помощи статистических инструментов нужно сделать вывод о подтверждении (или нет) гипотезы
- Вывод и рекомендации. Исследование должно быть ценно не само по себе, а из-за призыва к действию, полученного в результате
- Перспективы развития. Результаты исследования должны быть критически оценены, подсвечены ограничения с вариантами улучшения 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 отзывы, пользователей и рестораны (каждое из этих исследований будет иметь свои особенности). На первом этапе обсудите вопрос, на который вы хотите получить ответ с использованием данных.

   a. Например, при исследовании отзывов нас могут интересовать оценки, которые пользователи ставят ресторанам.

   b. При исследовании ресторанов можно сфокусироваться на среднем рейтинге.

   c. Если вы изучаете пользователей, то будет интересно смотреть на набор (или распределение) оценок, которые этот пользователь оставляет ресторанам (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.

3. 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 отзывом, оставленным пользователем о ресторане. Это дает возможность рассматривать в качестве единицы наблюдения и отзыв, и посетителя, и ресторан (но для этого нужны будут некоторые преобразования).

4. **Разведывательный анализ.** Проиллюстрируйте с помощью графиков распределения двух переменных, на которых вы решили сфокусироваться. Посчитайте основные описательные статистики для них. Какие особенности вы наблюдаете? Как их можно проинтерпретировать, т. е. объяснить?

   Примечание: если вам сложно сразу сформулировать исследовательский вопрос, можете начать работу над заданием с пункта 4, чтобы разобраться с предоставленным датасетом.

5. **Анализ взаимосвязей.** Посмотрев на переменные отдельно, можно переходить к проверке гипотезы. Отразите графически взаимосвязь между выбранными переменными 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.

6. Проверьте, являются ли результаты устойчивыми при анализе по подгруппам? Например, получаются ли похожие результаты по Москве и по Санкт-Петербургу? Схожи ли результаты для разных рубрик организаций (кафе, ресторанов и т. п.) или для разных групп посетителей? Проанализируйте взаимосвязь из предыдущего пункта, поделив выборку по какому-либо фактору. Объясните важность этого фактора. Как можно проинтерпретировать полученные результаты?

7. **Выводы и ограничения.** Для кого ваш результат может представлять интерес? 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, из тех, что могут содержаться в других источниках или среди факторов, которые сложно объективно измерить)? Что можно сказать о выборке пользователей/ресторанов, о ее репрезентативности? Можно ли распространять полученные выводы на все рестораны (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 механизм для обратного влияния.