Долгожданный Кекс

Постановка задачи

«Волшебная лавка мистера Оливандера» — крупнейший онлайн-магазин волшебной атрибутики, недоступной для Маглов. Её владелец, мистер Оливандер, недавно услышал о том, какую магию с его онлайн-ритейлом могут сотворить маркетологи, обученные анализу данных. Именно поэтому он пришёл за помощью к вам.

Мистер Оливандер довольно продвинутый маг. Он знает, что жизненный цикл клиента удобно изображать с помощью «воронки», описывающей движение человека по процессу покупки. Он понимает, что иногда бывает так, что клиенты приуныли. Обычно это происходит при переходе с одного шага воронки на другой.

Например, из 100 человек, увидевших рекламу, только 20 кликнут на неё и перейдут на сайт. Из перешедших на сайт только половина добавит товар в корзину, из добавивших товар в корзину, только четверть совершит покупку. На каждом этапе по каким-то причинам клиенты отваливаются. И это не очень круто. Увеличивая количество клиентов на каком-либо уровне воронки, мы автоматически увеличиваем результирующий объем продаж. Для этого надо как-то воздействовать на клиентов, но как именно непонятно.

Мистер Оливандер очень хочет избежать состояния клиентов «что-то приуныли» и повысить ключевые показатели своего бизнеса. Что за ключевые показатели, он толком не определился и надеется на ваши дельные советы. Вам необходимо сделать следущие вещи:

1. Определить как выглядит воронка магического онлайн-ритейла и понять где именно при проходе по ней могут отваливаться клиенты.
2. Определиться, на какие ключевые показатели, то есть метрики, необходимо ориентироваться на каждом этапе воронки, чтобы понимать улучшилась ситуация или ухудшилась. Для каждого этапа сформулируйте пул из нескольких метрик.
3. Понять как с помощью машинного обучения можно «улучшить» прохождение человека по воронке, определиться какие данные для этого использовать, откуда их взять.

Правила игры и оценивание

Кекс делается в группах по три человека. За сделанную работу группе выставляется 30 баллов. Баллы нужно поделить между членами группы самостоятельно.

Например, в группе работали Гермиона, Гарри и Рон. В итоге они получили 25 баллов. Гермиона работала больше всех, команда решает отдать ей 15 баллов. Гарри работал похуже, он получает 10 баллов. Рон — балбес, который ничего, по мнению команды, не делал. Он баллы не получает.

Ваша итоговая оценка зависит от того, насколько глубоко проработана задача, насколько чёт-

ко обоснован каждый этап работы, каждая метрика. Оцениваться будут следущие пойнты:

- Продуманность воронки, обоснованность каждой её части.
- Продуманность и обоснованность системы метрик для монтиринга прохода по воронке.
- Адекватность данных, выбранных для оценивания модели.
- Обоснование выбранных переменных. Чётко объясните почему вы выбрали в качестве целевой переменной именно то, что вы выбрали. Постарайтесь объяснить как именно ваша модель поможет достичь желаемого результата.

Обратите внимание, что от вас никто не требует описывать как вы делаете кросс-валидацию, дробите выборку и как вообще работает метод ближайшего соседа. Подразумевается, что это очевидно. Нас интересует именно ваше понимание того, как состыкуются между собой бизнес и машинное обучение, а не знание алгоритмов.

Куда сдавать и когда сдавать

На кекс у вас есть текущий семинар и ещё пара дней. Дедлайн - ??? Объём - ??? Две-три странички ворда без воды - ???

Дополнительная инфа: как ритейл собирает данные о нас

Для сбора данных и аналитики сайты используют специальные сервисы. Например, Яндекс.Метрику и Google Analytics. Эти сервисы позволяют анализировать то, что люди делают на сайте, с каких страниц они приходят, откуда они приходят (из поиска, с конкретного рекламного баннера и тп), какими демографическими характеристиками они обладают (пол, возраст, география и тп).

Более того, для каждого пришедшего пользователя существуют очень разношёрстные данные о визитах: в какой последовательности он смотрел страницы, куда кликал мышью, как ей двигал и т.п. Сервисы фактически показывает полную информацию о том, что происходит на сайте, позволяют выгрузить сырые данные и на их основе обучить какие-то модели.

Предположим, что Невил два года назад зашёл на сайт к мистеру Оливандеру. Вчера он сделал это повторно. Как понять, что эти два захода принадлежат одному и тому же человеку? Обычно для этого используют систему из разных id.

Если Невил заходил на сайт, используя свой личный кабинет, мы поймём, что это один и тот же человек по его внутреннему id. Другой способ идентифицировать человека — использовать его аккаунт в google или яндексе. Если человек зашёл на сайт, был залогинен в своей почте, и на сайте стояла метрика, мы сможем его отследить.

Более слабым идентификатором является device-id устройства, с которого работает человек. Ясное дело, что сначала человек может зайти с телефона, потом с компьютера и, если он не залогинен, то система будет думать, что это два разных человека.

Самым слабым идетификатором является id, построенный на основе куки человека. Если вы почистите в браузере куки, то этот id перезатрётся и система будет думать, что вы новый человек. Используя такую вложенную систему из адишников, мы можем понимать где выполнял действия один и тот же человек. Более подробно про метрику и несколько кексов, связанных с ней, рекомендую посмотреть интересную 30-минутную лекцию: https://events.yandex.ru/lib/talks/6063/. Возможно, вы почерпнёте из неё какие-то идеи для своего решения.

В оффлайн-ритейле дело обстоит немного сложнее. Когда у магазина есть два чека в базе данных, он никак не может понять принадлежат они одному и тому же человеку или нет. Чтобы как-то исправить эту ситуацию и научиться агрегировать покупки, магазины придумывают всякие ухищрения, позволяющие им накопить данные. Например, систему бонусных карт, по id которых можно понять, что чеки принадлежат одному и тому же человеку.