**Часть 3. Теоретическая**

*Основные бизнес-отчеты по работе онлайн-кинотеатра:*

- по просмотрам:
  количество фильмов, просмотренных и оцененных самым активным зрителем;
  top-5 наиболее активных пользователей за последние 3 месяца;
  в какие дни (условно, естественно, поскольку у нас имеются данные только по оценкам пользователей) наш кинотеатр был наиболее востребован у зрителей;

- зрительская аудитория нашего кинотеатра:
  возрастная структура зрительской аудитории;
  распределение по роду занятий;
  наиболее предпочитаемые жанры различными категориями пользователей;
  отток пользователей (количество пользователей, которые прекратили размещать свои оценки значительно раньше, чем завершился сбор данных по нашему кинотеатру) и какие фильмы привели к отказу от просмотров в нем;

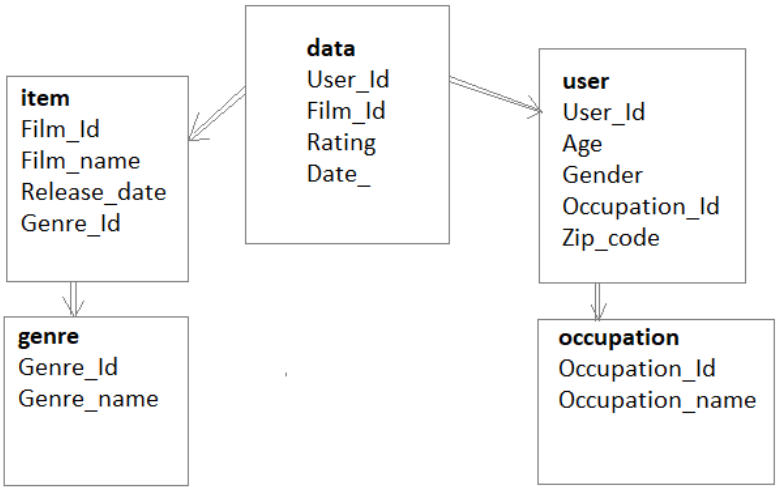
- по фильмам:
  фильмы, собравшие наибольшее количество оценок и наивысшую оценку пользователей;
  фильмы с самой широкой зрительской аудиторией (самый больший охват зрительских категорий).

*Основные данные и источники их поступления:*

Данные собираются с веб-сайта и очищаются. Критерием отбора данных является относительно постоянная активность пользователя  (каждый пользователь не может иметь менее 20 оценок к различным фильмам), кроме того данные должны быть полными и не иметь пропусков в аспекте демографической информации.

Датасет состоит из 100000 оценок 943 пользователей к 1682 фильмам. Имеются данные о фильмах, включающие идентификатор и название, дату выхода фильма и его жанровую принадлежность. Демографические данные о пользователях содержат идентификатор пользователя, его возраст, пол, род занятий, а также индекс.
Имеются справочная информация о жанрах представленных фильмов и список профессий.
Файлы  с первичными  данными конвертируются в csv-файлы, а те, в свою очередь являются источниками для создания хранилища данных и аналитических вычислений.

*Хранилище данных и процесс заливки данных:*




Существующие данные могут быть организованы в хранилище данных (схема звезда), которое включает основную таблицу - таблицу фактов **data**, ссылающуюся на справочники **item** и **user** со своими подсправочниками **genre** и **occupation** (соответственно). Таблица **data** содержит в себе данные о просмотрах в кинотеатрах, отражаемых с помощью суррогатного первичного ключа по сочетанию полей User_Id и Film_Id, и информации об оценках Rating, которые пользователи выставили по завершении каждого просмотра. Справочник **item** конкретизирует информацию по фильмам (имея данные о названии, годе выхода и принадлежности к конкретному жанру). Список жанров прилагается в подсправочнике **genre**. Справочник **user** содержит демографическую информацию о зрителях и дополняется расшифровкой их рода занятий в подсправочнике **occupation**. Следовательно, мы можем определить структуру хранилища как DDS DWH с требованиями соблюдения процесса ETL при загрузке данных, что предполагает определенную процедуру загрузки с четкой последовательностью этапов:

1. извлечения данных из источника;
2. очистки данных от ошибок;
3. приведения к одному виду;
4. приведения в соответствие с значениями из контекстных справочников;
5. загрузки подготовленных данных в хранилище.

Все эти требования обусловливают необходимость проведения проверки качества загружаемых данных.

*Основные проверки на качество данных:*

- в таблице фактов **data** поле Date_ должно включать значение даты с 19 сентября 1997 года по 22 апреля 1998 года;
- необходимо проверить являются ли вносимые данные валидными, например, 30 февраля считать невалидной датой;
- в поле Release_date справочника **item** значение даты должно умещаться в диапазон между 1895 и 1998 годом;

- в поле Age справочника **user** возраст пользователя - это целое число между 0 и 122;
- поле Gender справочника **user** включает только одно из двух возможных значений: M и F;
- в подсправочнике **occupation** значения поля Occupation_name должны быть уникальными;
- в справочнике **user** значение поля Occupation_Id должно однозначно ссылаться на поле Occupation_Id справочника **occupation**;
- значения поля Film_Id в справочнике **item** и поля User_Id в справочнике **user** представляют тип данных integer (Serial - в PostgreSQL) от 0 до 1681 и от 0 до 942 соответственно;
- в таблице **data** соответствие между значениями столбцов User_Id и Film_Id должно быть уникальным;
- значения поля Rating таблицы фактов **data** - это integer число от 1 до 5.

Кроме того, DDS DWH структура хранилища данных связана с регулярностью загрузки и обновляемостью данных.

*Новый Data-project и потребность в профильных специалистах*

Одна из основных целей в онлайн-кинобизнесе - увеличить количество просмотров контента, которую можно достичь как расширением аудитории пользователей видеосервиса, так и удержанием зрительского внимания со стороны постоянных клиентов. Для этой второй задачи и предлагается разработать рекомендательную методику их последующих просмотров и систему мониторинга результатов применения этой методики.
Разрабатываемая система рекомендаций для зрителей должна быть призвана предложить зрителю подходящий для его предпочтений контент, заинтересовать и убедить его в том, что он искал именно это. Рекомендации для пользователей должны быть персонализированными и основанными на изучении пользовательского опыта. Этот опыт будет полностью контролироваться алгоритмами.

Этапы и виды работ по проекту в соответствии с методологией CrispDm:

**Business Understanding**

- Можно предположить, что наш кинотеатр не сумел адекватно ответить на те вызовы, которые поставила перед отраслью видеосервисов современная глобальная экономика. И для того, чтобы полностью не утратить возможность генерировать денежный поток и не потерять клиентов, руководству онлайн-кинотеатра требуется найти прорывные решения в бизнесе. Одним из таких решений, предполагается, станет внедрение data-driven подхода, начало чему должна положить разработка персональной рекомендательной системы. Цели этого проекта - увеличение выручки онлайн-кинотеатра за счет увеличения просмотров, удержание клиентов и расширение зрительской аудитории.
- На этапе сбора исходной информации и оценки ситуации, в частности, необходимо определить механизм работы нашего кинотеатра, какие платформы

предлагаются зрителю, как осуществляется контакт со зрителем, какие модели монетизации использует наш кинотеатр. Какой экономический результат ожидается руководством от реализации идеи о внедрении персональной рекомендательной системы?

- В анализе ресурсов определяется персонал проекта, какие специалисты будут востребованы в ходе его разработки и внедрения (Аналитик, DS-специалист, Разработчик, Владелец продукта), какие данные будут использоваться для анализа (в нашем случае имеется исторический набор данных о завершенных просмотрах и оценках, которые выставили фильмам зрители, демографическая информация о зрителях и справочник профессий, данные о фильмах, включающие принадлежность каждого из них тому или иному жанру, год выхода и ссылку на страницу фильма в DMbd и справочник жанров). Мы предполагаем, что в силу обезличенности зрительской аудитории, нейтральности киноведческой информации и историчности просмотровых данных, проекту не сулят значительные риски юридического или экономического характера, и рентабельность проекта в этих условиях может быть вполне высокой. Интуитивно же мы полагаем, что эффект от внедрения более глубокого data-driven подхода, основанного на фиксации данных в реальном времени, будет как и более масштабным,так и более затратным, что потребует детального  анализа возможных выгод и затрат.
- Выводя цели и критерии успешного проведения проекта, необходимо определить как руководство компании представляет себе использование результатов и оценить желаемую точность будущей модели.
- Составление плана проекта.

Для успешного выполнения этого этапа необходимы **Аналитик** и **Владелец продукта**.

В конце первого этапа мы выполняем первоначальную оценку инструментов и методов. Выбирая алгоритм анализа данных, мы разделяем общую задачу - построение системы персональных рекомендаций - на две задачи меньшего масштаба: первая задача - кластеризация фильмов с применением алгоритма DBSCAN для разведочного анализа данных и предварительной обработки, а вторая задача - классификация фильмов в качестве рекомендуемых с помощью алгоритма логистической регрессии, выбранных из кластера наиболее интересных фильмов для пользователя.

Результатом всего этапа предположение, насколько выбранный нами инструментарий анализа данных окажется адекватным для достижения поставленной руководством цели и достаточны ли ожидаемые результаты для решения проблем компании.

На этапе **Data Understanding** мы производим:

- сбор данных, указанных в ресурсах проекта и их первоначальный анализ: описание данных - данные о просмотрах (в количестве 100000) фильмов даются уникальным сочетанием полей, содержащих идентификатор пользователя и идентификатор фильма, и дополняются оценкой пользователя и временем ее выставления; данные о пользователе (пол, возраст, род занятий),

данные о фильмах (идентификатор, название, дата выхода, принадлежность к жанру), справочники;
- исследование данных (традиционный анализ данных с использованием методов запросов, визуализации и отчетности, возможно рассмотреть несложные распределения ключевых признаков изучаемых сущностей, корреляции признаков, простые агрегации и статистический анализ);
- и определение их качества. Для проверки качества данных необходимы ответы на вопросы: достаточная ли полнота данных, охватывают ли они все кейсы, требующие решения, каковы ошибки в данных, насколько они распространены, как компенсировать недостающие значения в данных? Конкретно в нашем случае можно сказать, что данные подготовлены и не нуждаются в очистке.

Для успешного выполнения этого этапа необходим **Аналитик**.

В результате проведения работ необходимо добиться, чтобы все источники данных были четко определены и доступны, определить все проблемы с данными, выделить ключевые признаки и атрибуты.

На этапе **Data preparation** совершаем:
- выбор данных**,** разделение на обучающую и тестовую выборки, уточнение признаков, по которым будут анализироваться данные (пол, возраст, род занятий для пользователей, год выпуска, принадлежность к конкретному жанру для фильма, оценки фильмов пользователями);
- сохранение данных в data frame для обучения модели.

Для успешного выполнения этого этапа необходим **Аналитик**.

Условием перехода к следующему этапу являются возможность выбора подмножества для моделирования, получение качественных очищенных данных, составление релевантного набора данных, отражение результатов всех предыдущих этапов в документации.

На этапе **Modeling** предстоит:

- окончательно определиться с методом моделирования (убеждаемся, что для более точного предсказания предпочтений будущих просмотров и выработки рекомендаций для зрителя наиболее оптимальной является двухуровневая модель, включающая этап кластеризации методом DBSCAN с последующей классификацией фильмов методом логистической регрессии
- необходимо разработать альтернативные варианты построения модели;
- построить модели;
- оценить модель (ответ на вопрос, в какой степени результат применения модели соответствует желаемым критериям качества, какая из моделей показывает более высокое качество);
- описать результат.

Для успешного выполнения этого этапа необходим **DS-специалист**.

Результаты этапа **Modeling** должны соответствовать следующим критериям:
быть достоверными;
быть адекватно интерпретируемыми и релевантными для потребностей руководства и ведения бизнеса;
представить лучшее качество из нескольких разработанных моделей.

На этапе **Evaluation** необходимо оценить ценность модели анализа данных для бизнеса:

- насколько четко представлены результаты;
- позволяют ли результаты открыть новые инсайты;
- применимость к бизнесу;
- существуют ли какие-либо несовершенства, недоработки и ошибки;
- какие альтернативные действия могли бы быть выполненными;
- и запланировать дальнейшие действия.

Для успешного выполнения этого этапа необходим **Аналитик, DS-специалист** и **Владелец продукта**.

Этап **Deployment** предусматривает следующие шаги:

- планирование внедрения;
- планирование мониторинга и технического обслуживания;
- определение моделей и результатов, которые требуют поддержки;
- определения критериев понимания того, что модель перестала быть актуальной? Что делать в этом случае?
- итоговый обзор проекта.

Для успешного выполнения этого этапа необходим **Владелец продукта, Аналитик и Разработчик.**

*Полезные ссылки*:
https://www.the-modeling-agency.com/crisp-dm.pdf
https://vc.ru/marketing/31353-personalizaciya-oblozhek-na-netflix-mashinnoe-obuchenie-i-mnogorukie-bandity