## **Бизнес-отчеты**

Основной источник информации - это датасет, который содержит информацию о клиентах кинотеатра. Бизнес-отчет должен содержать гистограмму посетителей кинотеатра, которых я далее буду называть пользователями, по возрасту и по профессиям, в зависимости от их пола.

Для второго бизнес-отчета необходимо посмотреть на данные по фильмам, а именно важно видеть график количества фильмов по жанрам и количество фильмов по годам.

Для третьего бизнес-отчета нам нужно найти самых активных пользователей и вывести диаграмму количества оценок по месяцам и годам. Вывести в график 5 самых активных пользователей за последние 5 месяцев.

## **Основные имеющиеся данные**

Две таблицы. Первая Movies содержит id фильмов, все названия фильмов, год выпуска и отношение фильма к тому или иному жанру. Вторая основная таблица Ratings, содержит id пользователя, id фильма к которому пользователь оставил свой рейтинг (оценку).

## **Основные сущности**

За основные сущности берутся названия фильмов и id фильмов из первой таблицы, id пользователей и рейтинг просмотренного пользователем фильма из второй таблицы.

## **Основные проверки**

Для таблицы Movies - Id фильма (movie id) должен быть уникальным значением, как и название фильма. Дата выхода фильма должна быть не позднее 1997 года и содержать месяц и день. Ссылка на страницу фильма на сайте IMDb должна быть валидной, то есть открываться и иметь соответствующий названию фильм на этой странице. Значения для жанров фильмов должны содержать исключительно значения вида 0 или 1.

Для таблицы Ratings - значения id должны быть исключительно числовыми, как все остальные значения в таблицу Ratings. Значения столбца rating должны быть строго в диапазоне от 0 до 5. Значение timestamp может быть отформатированным в значение даты как для таблицы Movies. Movie id и item id подразумеваются как ключевые значения (связаны друг с другом).

## **Data-проект**

Целью дата-проекта стоит создание рекомендательной системы с точностью в 15% на основе данных, которые предоставляет наш дата-сет. Если проект окажется успешным и рекомендательная система будет работать с точностью в 15%, то показатель Customer retention (способность компании или продукта удерживать потребителей на протяжении определённого периода) возрастет, что в свою очередь повлияет на прибыль сервиса.

* Понимание бизнес-целей (Business Understanding).

В нашем случае мы должны понять цель по повышению точности рекомендательной системы. Необходимо уточнить, есть ли смысл в реализации проекта, насколько доход от рекомендательной системы превысит затраты на ее создание и поддержание в дальнейшем. Проверить хватает ли для создания проекта ресурсов что мы имеем, хватает ли штаба.

* Понимание данных (Data Understanding).

Необходимо понять, хватает ли данных для реализации нашего проекта, валидны ли эти данные. Найти связь одних данных с другими. Найти дополнительные источники для дополнения информации в случае их нехватки на этапе моделирования или подготовки данных.

* Подготовка данных (Data Preparation).

Обработка данных, их фильтрация, отбор всех необходимых данных. У нас это все id по фильмам и клиентам, названия фильмов, жанры по каждому фильму и просмотры от пользователей. В первую очередь нам необходимо заполнить все пропуски в данных, отредактировать данные содержащие лишние символы или не соответствующие стандартам остальным данных. Отформатировать данные для дальнейшей обработки, числовые значения - числовые, даты - даты.

* Моделирование (Modeling).

Выбираем модель в зависимости от задачи, в нашем случае подходит линейная регрессия. Создаем выборку, разделяем ее на тестовую и тренировочную. Обучаем модель на данных и получаем результат для обеих групп, сравниваем их результат.

* Оценка (Evaluation).

Проверяем, если наши данные выводят необходимые для нас 15% точности и полностью соответствуют нашей бизнес цели. На этом этапе мы принимаем решение о внедрении нашей бизнес-проекта

* Внедрение (Deployment).

Внедрение включает в себя сроки внедрения нашей рекомендательной системы в сайт и разработку этой самой системы.

## **Требуемые роли**

1. Руководитель проекта
2. Аналитик данных
3. Аналитик данных
4. Бизнес Аналитик
5. Data Scientist
6. Разработчик front-end
7. Разработчик back-end