Декабрь 2022г.

ИТОГОВАЯ РАБОТА

курс «Big Data с нуля»

Белов Константин
DAU-40

Содержание

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1. Задание №1. Описать основные бизнес-отчеты (2-3 штуки), которые мы хотим видеть по нашему бизнесу.

1.1. Отчет по продажам

Данный отчет включает в себя:

- Количество аренд фильмов по каждому месяцу.
- Сумма аренд фильмов по каждому месяцу.
- ТОП-50 фильмов по количеству просмотров за каждый месяц с указанием суммарной стоимости аренды этих фильмов за каждый месяц и стоимости их размещения.
- ТОП 50 фильмов по суммарной стоимости аренды за каждый месяц с указанием количества просмотров этих фильмов за каждый месяц и стоимости их размещения.
- ТОП 3 самых популярных жанра – по количеству просмотров и суммарной стоимости аренды за каждый месяц.

1.2. Отчет по пользователям.

Данный отчет представляет из себя сегментирование пользователей по группам. Каждая группа будет определяться: полом (Male, Female), возрастом (0-15, 16-20, 21-34, 35-49, 50-65, > 66 лет), регионом проживания. Для каждой группы за каждый месяц посчитано:

- Количество пользователей.
- Количество аренд.
- Частота использования сервиса.
- ТОП 3 любимых и ТОП 3 нелюбимых жанра.
- Сумма аренд фильмов.

1.3.Отчет по рекламе.

Данный отчет представляет из себя сводные данные по проведенным рекламным кампаниям:

- ID кампании.
- Наименование канала привлечение.
- Срок проведения кампании (в днях).
- Количество привлеченных пользователей.
- Стоимость кампании.
- Стоимость привлечения одного пользователя в рамках рекламной компании (учитываются только те пользователи, которые зарегистрировались на сайте/скачали приложение и арендовали хотя бы один фильм).

1.4. Общий отчет о делах онлайн-кинотеатра по каждому месяцу.

Данный отчет содержит в себе информацию:

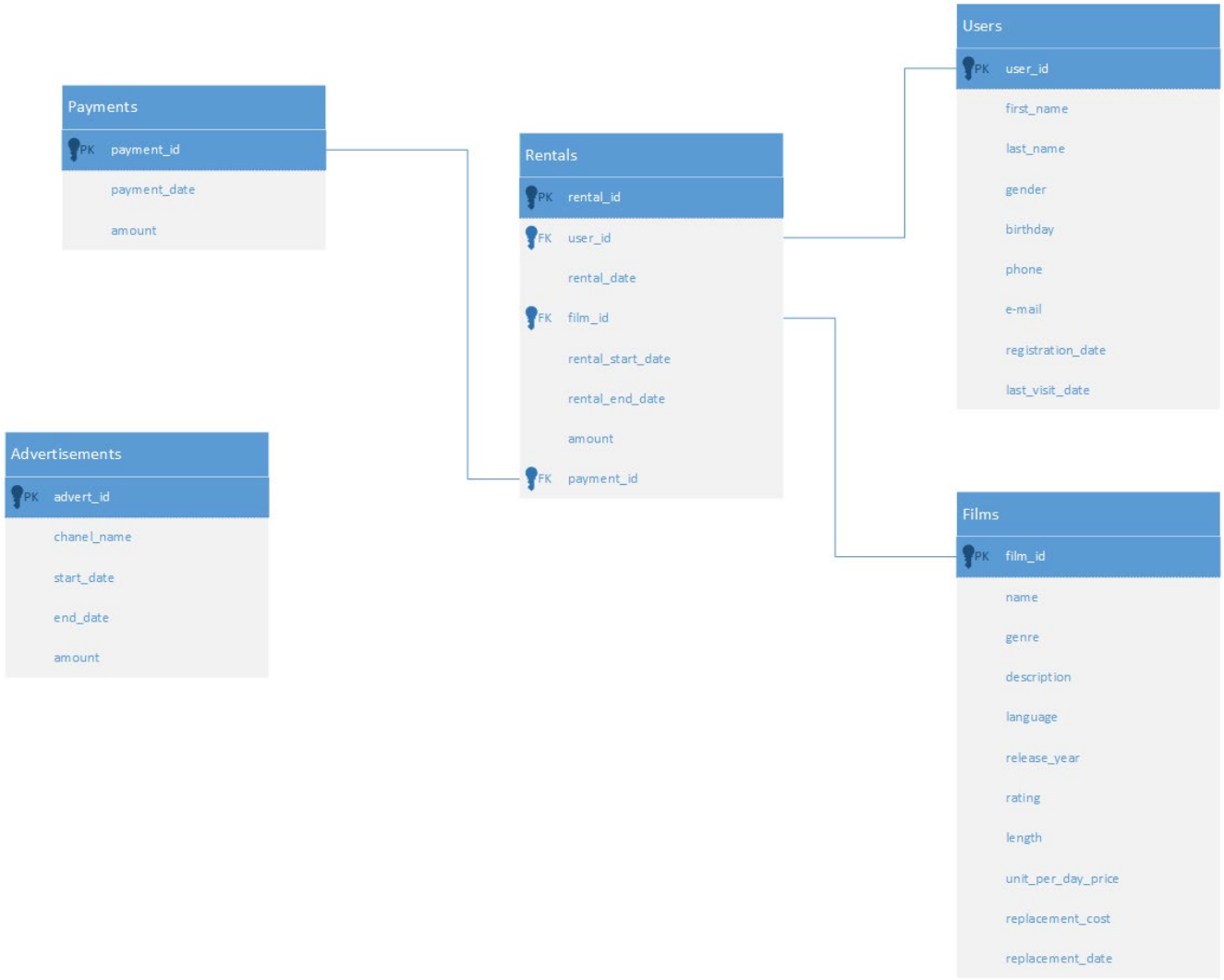
- Выручка за каждый месяц.
- Расходы на размещение фильмов.
- Расходы на аренду офиса, серверов в ЦОДе, персонал и пр.
- Расходы на рекламу.
- Прибыль за каждый месяц.
- Итоговая выручка за год.
- Итоговые расходы за год.
- Итоговая прибыль за год.

2. Задание №2. Описать основные имеющиеся данные и источники их поступления.

2.1 Информация о пользователях интернет кинотеатра. Источник данных: сайт и мобильное приложение
2.2 Информация об арендах фильмов. Источник данных – сайт, мобильное приложение, CRM система.
2.3 Инофрмация о произведенных платежах. Истоник данных: интернет-банк, 1С Бухгалтерия.
2.4 Инофрмация о размещенных фильмах. Источник данных – 1С Бухгалтерия (Управление торговлей, Управление запасами).
2.5 Информация о проведенных рекламных кампаниях. Источник данных: платформы по привлечению клиентов, 1С Бухгалтерия (в разрезе дат платежей).

3. Задание №3. Описание основных сущностей в хранилище данных (схема звезда) и процесса заливки данных.

3.1 Хранилище данных имеет следующую структуру:



3.2 Описание таблиц:

3.2.1 Таблица Advertisements

Данная таблица содержит информацию о рекламных кампаниях, направленных для привлечения пользователей.

Столбец	Тип данных	Модификаторы	Индексы	Описание
advert_id	int	NOT NULL	PRIMARY KEY	Уникальный номер рекламной кампании
chanel_name	text	NOT NULL		Название канала привлечения
start_date	date	NOT NULL		Дата начала кампании
end_date	date			Дата окончания кампании
amount	numeric			Стоимость кампании

Поля end_date и amount заполняются после окончания кампании.

3.2.2 Таблица Films

Данная таблица содержит информацию о фильмах кинотеатра.

Столбец	Тип данных	Модификаторы	Индексы	Описание
film_id	int	NOT NULL	PRIMARY KEY	Уникальный номер фильма

name	varchar(50)	NOT NULL		Название фильма
genre	varchar(20)	NOT NULL		Жанр
description	text			Описание фильма
language	varchar(50)	NOT NULL		Язык фильма
release_year	date	NOT NULL		Год выхода
rating	varchar(5)	NOT NULL		Возрастной рейтинг
length	time	NOT NULL		Продолжительность
unit_per_day_price	numeric	NOT NULL		Стоимость аренды в день
replacement_cost	numeric	NOT NULL		Стоимость размещения
replacement_date	date	NOT NULL		Дата размещения

3.2.3 Таблица Payments

Данная таблица содержит информацию об оплатах аренды.

Столбец	Тип данных	Модификаторы	Индексы	Описание
payment_id	int	NOT NULL	PRIMARY KEY	Уникальный номер платежа
payment_date	timestamp	NOT NULL		Дата платежа
amount	numeric	NOT NULL		Сумма платежа

3.2.4 Таблица Rentals

Данная таблица содержит информацию об арендах фильмов.

Столбец	Тип данных	Модификаторы	Индексы	Описание
rental_id	int	NOT NULL	PRIMARY KEY	Уникальный номер аренды
user_id	varchar(50)	NOT NULL	FOREIGN KEY (Users)	Номер пользователя
film_id	text	NOT NULL	FOREIGN KEY (Films)	Номер фильма
rental_start_date	date	NOT NULL		Дата начала аренды
rental_end_date	date	NOT NULL		Дата окончания аренды
amount	numeric	NOT NULL		Сумма аренды
payment_id	int	NOT NULL	FOREIGN KEY (Payments)	Номер платежа

*Поле amount является вычисляемым: (Rentals.rental_end_date – Rentals.rental_start_date) * Films. unit_per_day_price*

3.2.5 Таблица Users

Данная таблица содержит информацию о пользователях.

Столбец	Тип данных	Модификаторы	Индексы	Описание
user_id	int	NOT NULL	PRIMARY KEY	Уникальный номер пользователя
first_name	varchar(50)	NOT NULL		Имя пользователя
last_name	varchar(50)	NOT NULL		Фамилия Пользователя
gender	varchar (6)	NOT NULL		Пол
birthday	date	NOT NULL		Дата рождения
phone	varchar(11)	NOT NULL		Телефон
e-mail	varchar(30)	NOT NULL		Электронная почта
registration_date	date	NOT NULL		Дата регистрации
last_visit_date	date	NOT NULL		Дата последнего посещения

Загрузка данных в таблицы Payments, Rentals и Users происходят с сайта онлайн-кинотеатра (в т.ч. и из мобильного приложения).
Загрузка данных в таблицу Films происходит из ERP-системы.
Загрузка данных в таблицу Advertisements происходит из платформ по привлечению клиентов (Yandex Direct, Google Adds и т.п.).

4. Задание №4. Описать основные проверки на качество данных (10 штук), которыми будем пользоваться при заливке.

4.1 В таблицах **Advertisements, Films и Payments** поля

- Advertisements.amount,
- Payments.amount
- Films.unit_per_day_price
- Films.replacement_cost

должны быть заполнены и содержать числовое значение, округленное до 2 знаков после запятой.

4.2 В таблицах **Advertisements, Films, Rentals и Users** поля

- Advertisements.start_date
- Advertisements.end_date
- Films.replacement_date
- Films.release_year
- Rentals.rental_start_date
- Rentals.rental_end_date
- Users.registration_date
- Users.last_visit_date

Должны быть заполнены и содержать значения, типа “YYYY-MM-DD”, где YYYY – год, MM – месяц, DD – день.

4.3 В таблице **Users** поле

- Users.phone

должно быть заполнено у каждого пользователя и содержать числовое значение, формата: “+XXXXXXXXXXX”, где X – число от 0 до 9. Это должно быть валидное значение, уникальное в рамках системы.

4.4 В таблице **Users** поле

- Users.email

должно быть заполнено у каждого пользователя и содержать значение формата <text>@<text>.<text>. Это должно быть валидное значение, уникальное в рамках системы.

4.5 В таблице **Users** поля

- Users.first_name
- Users.last_name

должны быть заполнены у каждого пользователя. Допускается использовать кириллицу или буквы латинского алфавита, а так же следующие специальные символы: < >, < - >.

4.6 В таблице **Payments** поле

- Payments.payment_date

должно быть заполнено и содержать значение, типа “YYYY-MM-DD-hh-mm-ss”, где YYYY – год, MM – месяц, DD – день, hh – часы, mm – минуты, ss – секунды.

4.7 В таблице **Films** поля

- name
- genre

должны быть заполнены у каждого фильма. Допускается использовать кириллицу или буквы латинского алфавита, а так же следующие специальные символы: < >, < - >, < , >, < ! >, < ? >, < : >, < . >, < – > и цифры от 0 до 9.

4.8 В таблице **Films** поле

- language

должно быть заполнено у каждого фильма. Допускается использовать буквы латинского алфавита.

4.9 В таблице **Films** поле

- description

должно содержать текст произвольной длины. Допускается использовать кириллицу или буквы латинского алфавита, а так же следующие специальные символы: < >, < - >, < , >, < ! >, < ? >, < : >, < . >, < – > и цифры от 0 до 9.

4.10 В таблице **Films** поле

- length

должно быть заполнено у каждого фильма и содержать значения типа: hh-mm-ss, где hh – часы, mm – минуты, ss – секунды.

5. Задание №5. Придумать Data-проект, который должен улучшить показатели Вашего бизнеса и расписать его по Crisp-DM.

Совет директоров онлайн кинотеатра на основании отчетов о состоянии дел в компании, приняло решение о более агрессивном продвижении кинотеатра и захвату рынка.

Задача на следующий год: увеличить прибыль компании на 30% .

Для достижения этой цели, были выделены три направления:

1. Увеличить рекламный бюджет.
2. Ввести три типа подписки: 1 месяц, 6 месяцев и 12 месяцев. Все три типа будут одинаковы по наполнению (фильмам, которые будут входить в подписку). Ознакомившись с отчетами о просмотрах фильмов, руководство утвердило правило, по которому фильмы будут включаться в подписку: 20% подписки – фильмы, которые посмотрело 80% посетителей за последний год, 50% подписки – фильмы, которые просмотрели 50% посетителей за последний год, 30% подписки – фильмы, которые просмотрели 20% посетителей за последний год.
3. Увеличить количество фильмов, представленных в онлайн-кинотеатре.

Перед руководителем отдела Аналитики Советом директоров поставлены вопросы:

1. На сколько % должен быть увеличен рекламный бюджет.
2. Какие фильмы следует включить в подписку.
3. Какова должна быть стоимость каждого типа подписки для пользователя.
4. Фильмы из каких жанров и с каким рейтингом необходимо прежде всего разместить в онлайн-кинотеатре.
5. На сколько % от текущего количества необходимо увеличить количество фильмов.

Исходя из поставленных вопросов, руководителем отдела Аналитики были выделены следующие направления работы:

1. Произвести анализ предпочтений пользователей и выделить:

Фильмы, имеющие, которые посмотрело 80%, 50% и 30% пользователей.

Наиболее и наименее популярные жанры и возрастные рейтинги.

2. Построить модель, предсказывающую, на сколько % увеличится количество аренд фильмов текущими пользователями, при увеличении количества фильмов в каждом жанре. % увеличения количества фильмов: 20% (план минимум), 30% (план средний) и 50% (план максимум).
3. Построить модель, предсказывающую, на сколько % увеличится количество привлеченных пользователей при увеличении рекламного бюджета на 30% (план миниум), 50% (план средний) и 70% (план максимум)
4. На основании п.п. 2 и 3 спрогнозировать количество аренд фильмов по каждому жанру и предполагаемую сумму трат пользователя с учетом увеличения количества фильмов и количества пользователей.
5. Исходя из п 4. предложить варианты стоимости каждого типа подписки.
6. Исходя из п. 3-5 представить прогноз, на сколько увеличится прибыль компании в следующем году. Прогноз необходимо представить в трех вариантах: минималистичный, реалистичный и оптимистичный.

План проекта по Crisp-DM

Этап	Задачи	Роли
Business Understanding	1. Вывести метрику определения популярности для каждого фильма и жанра. Определить типичные значения. 2. Вывести метрику стоимости привлечения пользователя по каждому каналу продвижения. Определить типичные значения, проверить наличие сезонности. 3. Провести когортный анализ пользователей (по дате регистрации), вывести Retention Rate, Churn Rate. 4. Провести когортный анализ пользователей по предпочтениям, вывести LTV по каждой когорте. 5. Для оценки качества созданных моделей посчитать: • % увеличения количества аренд в текущем году по сравнению с предыдущим • % увеличения количества фильмов в текущем году, по сравнению с предыдущим • % увеличения рекламного бюджета в текущем году по сравнению с предыдущим 6. Определить, как заказчик видит использование полученных моделей, сформулировать минимально необходимое качество. 7. Оценить ожидаемый эффект от моделирования и сравнить его с ожидаемыми трудозатратами.	Владелец продукта, аналитик
Data Understanding	1. Собрать все имеющиеся данные воедино (информация о пользователях и история аренд, информация о жанрах, фильмах и возрастных рейтингах, информация о стоимости рекламных кампаний) 2. Описать собранные данные (количество, типы, значения, связи). 3. Провести оценочное исследование данных. 4. Оценить качество данных на предмет наличия ошибок, отсутствующих значений, полноты.	Аналитик
Data Preparation	1. Очистить данные от пропусков, повторяющихся значений и ошибок. 2. Обогатить данные для увеличения количества признаков, на основании которых будет предсказываться результат. 3. Разделить данные на выборки test и train 4. Сохранить выборки в отдельные датафоеймы.	DS, DE
Modeling	1. Выбор модели – поскольку во всех случаях мы предсказываем целевую переменную (количество аренд фильмов, количество привлеченных пользователей) – это будет модель линейной регрессии. 2. Построить модель линейной регрессии. 3. Произвести оценку качества модели. 4. Описать результат моделирования.	DS

Evaluation	1. Оценить результат. Для оценки используем сравнение значений предсказанных целевых переменных со значениями, полученными на этапе Business Understanding. 2. Если результат не соответствует ожиданиям – анализ процесса моделирования, поиск альтернативных решений (например, до обогащение данных за счет дополнительных источников) 3. Определить дальнейшие шаги (либо переход на этап Business/Data Understanding, подготовка в внедрению).	Аналитик
Deployment	1. Создание плана внедрения для каждой модели. 2. Определение возможных проблем при внедрении. Описать пути их решения. 3. Составить план мониторинга и технического обслуживания (при необходимости) 4. Провести итоговый обзор проекта.	Владелец продукта, Аналитик, DS, Разработчик

6. Задание №6. Описать требуемые роли в команде по работе с данными на этапах 4 и 5.

Описано в п. 5 задания.

ПРАКТИКА GOOGLE SHEETS

Задания по данной практике приложены отдельным файлом.

ПРАКТИКА PYTHON

Задания по данной практике приложены отдельным файлом.