Теоретическая часть

*Задача: разработать стратегию внедрения хранилища данных и работы с большими данными в онлайн-кинотеатре.*

**#1 Бизнес-отчеты**

1) Оценка притока новых клиентов и удержания старых.

Цель: Повышение точности планирования бизнес-показателей.

2) Оценка механизма рекомендаций в онлайн-кинотеатре.

Цель: Выработать решения по улучшению рекомендаций.

**#2 Данные и источники их поступления**

| **Данные** | **Источники** |
| --- | --- |
| Персональные данные зарегистрированных пользователей | CRM |
| Статистика нахождения на сайте | Трекер на сайте |
| Фильмы | База данных фильмов |
| Статистика по досмотру фильмов | Трекер на сайте |
| Оформление пользователями подписок | Интернет-эквайринг, принимающий платежи |
| Данные по поиску по сайту | Поисковый модуль |

**#3 Хранилище (схема звезда) и процесс заливки данных**

****

**#4 10 проверок на качество данных**

1) ID пользователя: уникальное сочетание цифр, целое положительное число

2) Электронная почта: сочетание латинских букв и цифр, нижнее подчеркивание, точка + обязательное наличие @ + существующий сервер + доменное расширение; нечувствительно к регистру

3) Пароль: не короче шести символов, сочетание латинских букв и цифр, допустимо использовать символы, чувствительно к регистру

4) Возраст: положительное целое число от 6 до 120

5) Телефон: номер в формате +X XXX XXX XX XX.

6) Timestamp: формат Unix

7) Рейтинг: положительное число от 0 до 5, выраженное в десятичных долях

8) Пробная подписка: булево значение true/false

9) Дата подписки: число от 01 до 31 + число от 01 до 12 + 4 цифры в пределе от 2000 до 2050

10) Имена актеров: не менее одного слова из одной и более кириллических букв

**#5 Улучшение бизнес-показателей (Crisp-DM)**

Цель проекта: увеличить время пребывания на сайте зарегистрированных пользователей

| **Бизнес-анализ** | **Анализ данных** | **Подготовка данных** | **Моделирование** | **Оценка решения** | **Внедрение** |
| --- | --- | --- | --- | --- | --- |
| Цель: увеличить время пребывания на сайте.  Текущая ситуация:  time avg = 1,2 hour    Цель аналитики: найти слабые места и методы улучшения ситуации    План проекта: на основе данных из хранилища найти популярные жанры и предсказать рейтинг, чтобы улучшить механизм рекомендаций | Сбор:  Обратиться к существующим базам    Описание:  Подготовить первичный анализ    Изучение:  Выявить метрики (время пребывания на сайте, bounce rate, life-time value)      Проверка:  Проверить валидность данных | Выборка:  Отобрать данные, установить связи    Очистка:  Удалить лишнее, восполнить пробелы, конвертировать    Генерация:  Сделать производные данные    Интеграция:  Объединить данные    Форматирование: привести в нужный формат | Выбор алгоритмов:  Кластеризация, регрессия    Подготовка плана тестирования    Обучение моделей      Оценка их качества | Оценка результатов: сплит-тесты    Оценка процесса: подсчет метрик      Оценка следующих ходов: рекомендации бизнесу | Внедрение:  Запланировать развертывание нового механизма рекомендаций    Планирование мониторинга: найти команду поддержки и обучить      Подготовка отчета: сделать финальный отчет    Ревью: оценить проект и сформулировать next steps |

**#6 Роли в команде**

4. Modeling: **Data scientist** строит и валидирует модели, помогает решать задачи предсказания

5. Evaluation: **Аналитик данных** оценивает решения, проводит тестирование, дает рекомендации бизнесу