# Теоретическая часть

Вы - главный по данным в среднем по объему просмотров интернет-кинотеатре. Ваша задача разработать стратегию внедрения хранилища данных и работы с большими данными в этой компании. Задания:

**1**.Описать основные бизнес-отчеты (2-3 штуки), которые мы хотим видеть по нашему бизнесу.

* отчет об оттоке старых клиентов(сколько клиентов не стало продять подписку)(хочу видеть как много старых клиентов отказывается от нашего сервиса и по каким причинам)
* отчет о привлечении новых клиентов (хочу знать как много и по каким каналам к нам приходят новые клиенты)
* отчет об активности клиентов(хочу видеть как много времени наши клиенты тратят на сервис, как много подходящего контента для них у нас есть)

**2**.Описать основные имеющиеся данные и источники их поступления.

Основные имеющиеся у нас данные:

* данные пользователя , поступают от самих пользователей при регистрации
* данные о фильмах и сериалах, поступают при появлении контента на сервис
* данные о просмотрах и оценках, появляются при просмотрах пользователями фильмов
* данные о платной подписке, приобрел ли, продлил ли пользователь подписку

**3**.Описать основные сущности в хранилище данных (схема звезда) и процесс заливки данных.

Основные сущности:

* фильм или сериал
* пользователь
* подписка
* дата и время
* транзакция

**4**.Описать основные проверки на качество данных (10 штук), которыми будем пользоваться при заливке.

* имейл пользователя должен быть уникальным и желательно подтверждаться.
* телефон пользователя должен быть уникальным и возможно желательно тоже подтверждаться.
* имя, фамилия должны быть заполнены корректно
* данные о фильмах и сериалах должны быть заполнены и быть реальными
* проверка оплаты подписки клиентом
* обязательно дата и время и другие данные проведения транзакции
* у транзакции должен быть строго один пользователь , тут всё очень строго
* например проверка уникальности пользователя в уникальной подписке(одна подписка - один пользователь , если у нас такие правила, или не больше определенного количества человек)

**5**.Придумать Data-проект, который должен улучшить показатели Вашего бизнеса и расписать его по Crisp-DM.

Проект - создать модель машинного обучения предполагающую общую удовлетворенность клиента сервисом(отток клиентов). Если модель сможет выявить неудовлетворенность сервисом(желание уйти) то сможет предложить скидки или другие привилегии клиенту. Это поможет уменьшить отток клиентов и соответственно улучшить показатели бизнеса.

1.Business Understanding

Успех будет определяться уменьшением количества уходящих клиентов

Риски - модель не сработает.

Бизнес цель - уменьшить отток клиентов, повысить общую удовлетворительность сервисом клиентов и как следствие улучшить показатели.

2.Data Understanding

Сбор данных, хватает ли нам данных? Хорошего ли качества наши данные?

3.Data Preparation

Выборка, очистка, генерация, форматирование данных. Подготовка data frame для обучения модели и сохранение его.

4.Modeling

Выбор наиболее подходящего алгоритма. Построение модели. Оценка качества модели.

5.Evaluation

Оценка результата. Удовлетворительность результатами? Есть ли новые вводные данные?

6.Deployment

Планирование внедрения модели. Подготовка отчетов. Планирование мониторинга и обслуживания. Полный ревью проекта.

**6**.Описать требуемые роли в команде по работе с данными на этапах 4 и 5

На этапе 4 проверка качества думаю нужны в первую очередь разработчики, проверки должны быть на уровне кода, а так же аналитики которые укажут на необходимые проверки.

На этапе 5 подпункт 1 нужны аналитики, подпункт 2 - аналитики, 3 - DS, DE, 4 - DS, 5 - DS, аналитики, 6 - DS и разработчики.