## 1. Постановка задачи

Бизнес-постановка задачи и ML-задача:

Цель: снижение оттока клиентов в онлайн кинотеатре, который предоставляет доступ к фильмам по модели подписки.
Проблема: большое количество пользователей отменяют подписку через определённое время, что негативно сказывается на доходах компании.
Бизнес-задача: разработать систему предсказания оттока клиентов для онлайн-кинотеатра, которая поможет маркетинговой команде своевременно реагировать и разрабатывать программы лояльности.

Задача машинного обучения заключается в построении модели бинарной классификации, которая на основе исторических данных о поведении пользователей сможет предсказывать, откажется клиент от подписки или нет.

Набор данных:

- Демографические данные (возраст, пол, регион).
- История подписки (время подписки, время отмены).
- Потребительская активность (количество просмотров, жанры предпочитаемого контента).
- Информация о платежах (способ оплаты, наличие просрочек, использование промо-кодов).
- Обратная связь (оценки контента, жалобы, обращения в поддержку).

Пример набора взят из открытых источников. Набора данных: Kaggle: Telco Customer Churn Dataset (https://www.kaggle.com/datasets/blastchar/telco-customer-churn ), который можно адаптировать для стримингового сервиса, заменив телеком-услуги на контентные.

## 2. Выбор и обоснование метрики для измерения качества

Для задачи предсказания оттока клиентов у нас может быть дисбаланс классов: подавляющее большинство пользователей останутся на сервисе, и только малая доля клиентов уйдет. В таких случаях метрика Accuracy (точность) может быть недостаточно информативной, так как модель, предсказывающая "остается" для всех клиентов, будет иметь высокую точность, но не будет полезна.

Для нашей задачи наиболее подходящие метрики:

- Precision — доля правильных предсказаний от всех предсказаний "отток". Важно минимизировать ложные срабатывания (false positives), чтобы маркетинг не тратил ресурсы на тех, кто не собирается уходить.
- Recall — доля правильно предсказанных уходящих клиентов от всех реальных уходящих клиентов. Важно захватить как можно больше пользователей, которые действительно уйдут.

- F1-score — гармоническое среднее между Precision и Recall, которое балансирует между тем, чтобы находить максимальное количество уходящих клиентов и минимизировать ложные срабатывания.
- ROC-AUC (Area Under the Curve) — метрика, которая показывает, как хорошо модель различает классы. Она полезна для оценки классификаторов на разных уровнях порогов.

1. F1-score будет ключевой метрикой, так как она учитывает и Precision, и Recall, что важно при дисбалансе классов. Мы стремимся найти как можно больше клиентов, которые могут уйти, при этом минимизируя ложные тревоги.
2. ROC-AUC будет дополнительной метрикой для оценки общей способности модели различать классы, особенно на ранних этапах.

Выбор метрик:

- Основная метрика — F1-score.
- Дополнительная метрика — ROC-AUC.

## 3. Проведение EDA (Exploratory Data Analysis)

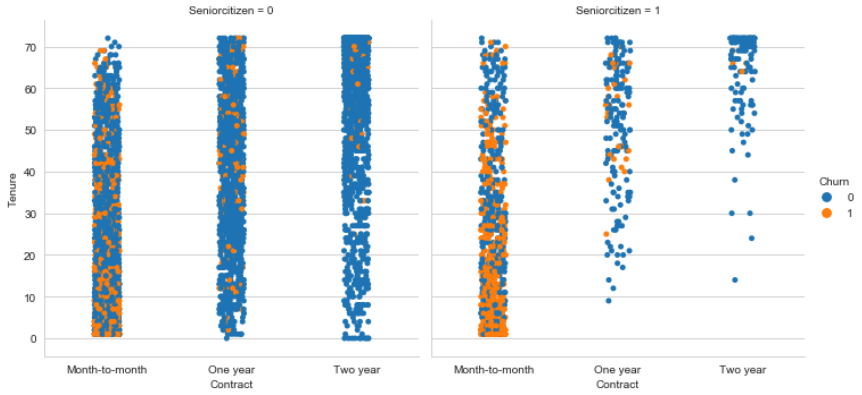
Для начала, посмотрим на типы и характеристики набора данных:

```
Shape of the data (7043, 21)
Feature names and data types ::
Customerid         object
Gender             object
Seniorcitizen       int64
Partner            object
Dependents         object
Tenure              int64
Phoneservice       object
Multiplelines      object
Internetservice    object
Onlinesecurity     object
Onlinebackup       object
Deviceprotection   object
Techsupport        object
Streamingtv        object
Streamingmovies    object
Contract           object
Paperlessbilling   object
Paymentmethod      object
Monthlycharges    float64
Totalcharges       object
Churn              object
```

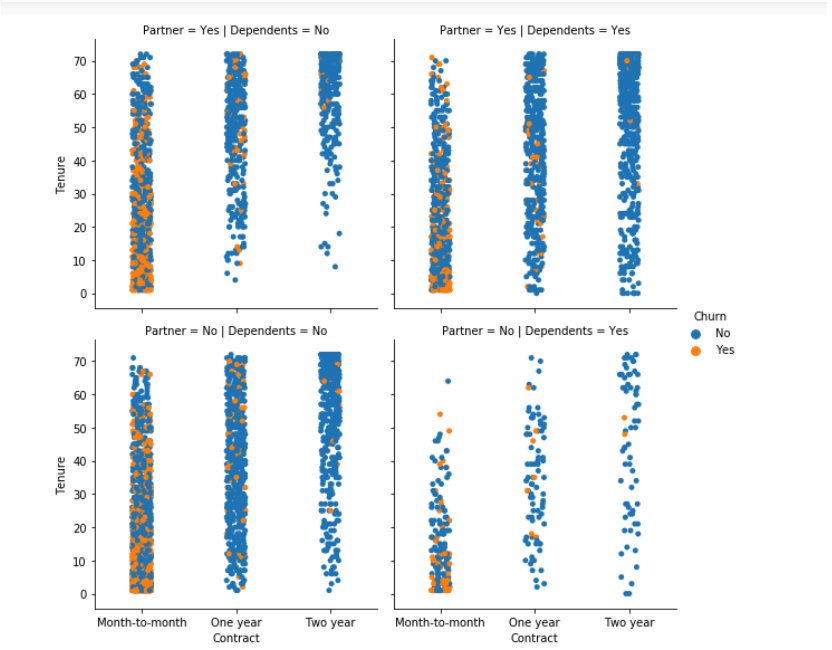
**Я буду рассматривать данные для телеком компании без интерпретации для онлайн кинотеатра**. Большинство признаков в наборе данных являются категориальными. Поэтому такие показатели, как среднее значение и медиана, могут применяться только к числовым признакам, например, к сроку пребывания и расходам.

Имеется четыре характеристики клиента, которые описывают его индивидуальный профиль: пол (мужчина/женщина), наличие партнера (да/нет), наличие иждивенцев (да/нет) и статус пенсионера (да/нет). Другие 16 признаков касаются предоставляемых услуг (Телефон/Интернет) и конкретных опций в этих услугах. Каждый клиент пользуется либо телефонной связью, либо интернет-услугами у нас.

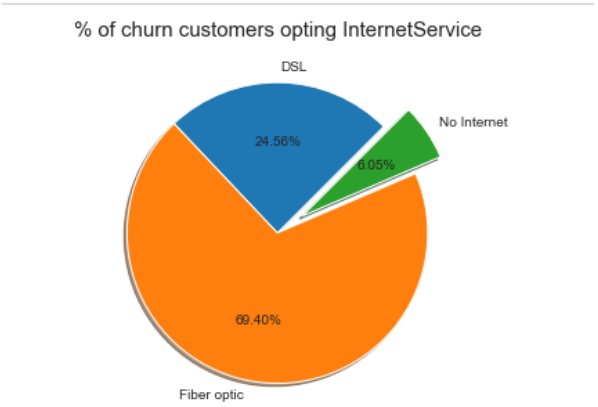
Срок действия услуг измеряется в месяцах: минимальный срок составляет 0 месяцев (для клиентов, которые недавно начали пользоваться услугами), а максимальный срок — 72 месяца (6 лет), что включает клиентов, пользующихся нашими услугами более 6 лет.




- За 6 лет общее количество клиентов составило 7043.
- 83% клиентов — это работающие/студенческие группы клиентов, 54% из которых имеют как телефон, так и интернет-услуги. В этой группе клиентов наблюдается отток в размере 25% за эти годы.
- Однако среди пенсионеров наблюдается большее отток, большинство из них находятся на ежемесячном плане обслуживания. Это может быть связано с финансовыми причинами.



- У 70% клиентов нет никаких зависимостей, а у 50% клиентов нет никаких партнеров. Так что может быть, что в основном есть клиенты с подписками на услуги только для себя.
- Клиенты с партнерами и/или иждивенцами предпочитают годовые подписки и меньше оттоков. Клиенты с ближайшими родственниками (рабочий класс) не предпочли бы постоянно менять оператора услуг.




- У нас на 12% больше подписок на телефонные услуги, чем на интернет-подписки. Мы можем выдвинуть гипотезу, что наши клиенты рассматривают нас как компанию телефонных услуг.
- Общий процент оттока составил 26,5%, при этом 93% этих клиентов имеют подключение к Интернету.
- 69% отошедших клиентов выбрали оптоволоконный сервис по сравнению с 24,5% отошедших клиентов с DSL-сервисами. Это может быть связано с отсутствием интернет-услуг с оптоволоконным сервисом из-за плохой инфраструктуры или плохой пропускной способности.