Урок 3. Создание проекта машинного обучения

* Подумайте о реальной проблеме, которую, по вашему мнению, можно решить с помощью машинного обучения. Четко определите эту проблему и поставьте SMART-цели для ее решения. Помните, что это должна быть проблема, в которой машинное обучение может помочь сделать прогноз, классифицировать данные или дать рекомендации.
* Определите потенциальные источники данных, которые могут быть использованы для решения вашей проблемы. Какие данные вам понадобятся? Где вы можете их получить? Напишите краткое резюме вашей стратегии сбора данных.
* Даже если у вас нет реальных данных для анализа, представьте, что они у вас есть. Как бы вы стали исследовать и понимать эти данные? Что бы вы искали? Напишите краткое описание вашей стратегии анализа исследовательских данных.
* Вспомните различные модели машинного обучения, которые мы обсуждали, в частности модель "Дерево решений". Какой тип модели машинного обучения будет наиболее подходящим для вашей проблемы и почему? Объясните свой выбор.

1. Напишите краткое описание того, как вы будете обучать и оценивать выбранную вами модель машинного обучения, используя воображаемый набор данных.
2. Наконец, подумайте о том, как вы могли бы развернуть свою обученную модель. Где она будет использоваться? Кто будет ее использовать? Какую пользу она принесет им? Напишите краткое резюме вашей стратегии развертывания.

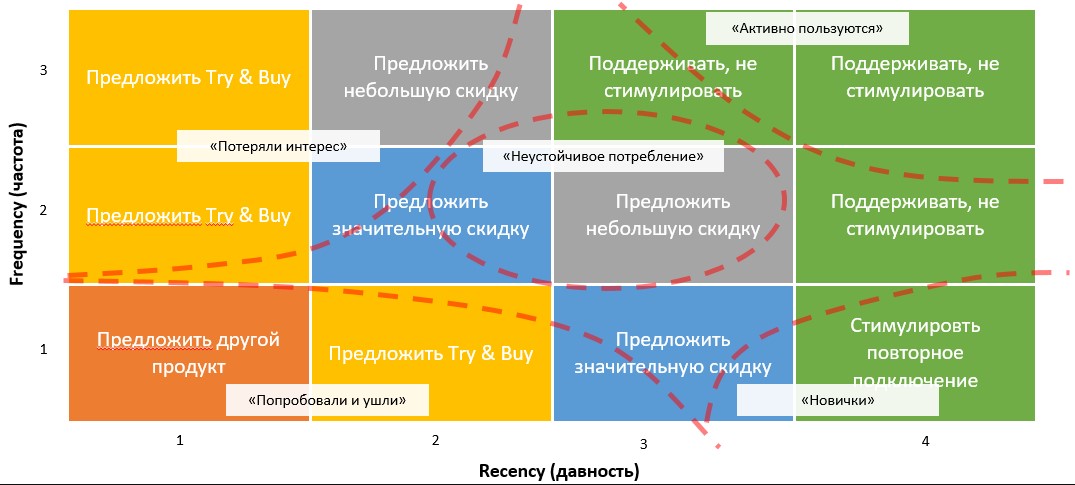
Помните, что целью этого задания является не создание реальной модели машинного обучения (пока!), а применение концепций, которые вы изучили на этом семинаре, к реальной проблеме.

**Задача сегментации**

В машинном обучении эффективное использование данных имеет первостепенное значение. Сегментация данных является важным процессом в этой сфере, обеспечивающим организацию и анализ наборов данных для получения инновационной информации. Сегментация данных играет ключевую роль в машинном обучении — от повышения точности моделей до оптимизации процессов принятия решений.

Сегментация клиентской базы — это способ повышения эффективности работы с пользователями путем их распределения по отдельным группам, или сегментам, в соответствии с их запросами и/или потребностями.

Иными словами, мы хотим понять, чем пользователи отличаются друг от друга



Примеры задач сегментации клиентских баз:

* интернет провайдеры
* мобильные операторы
* маркетплейсы (Ozone, Сбермарект и др.)
* банки и др.

Как правило, у вышеупомянутых организаций большое количество данных о клиентах (дата регистрации, количество покупок, возраст, пол и др.)

Для понимания данных, определим виды сегментации:

* по психотипам
* по интересам
* по платежам
* По эффективности взаимодействия
* Когда клиент совершил первое взаимодействие (сколь прошло времени)

Предварительная обработка данных и выбор алгоритмов

Сегментация данных без учителя — это метод машинного обучения, используемый для разделения данных на значимые и однородные группы или кластеры без предварительного знания меток или категорий. Этот подход особенно полезен при работе с большими наборами данных, когда вручную маркировать каждый экземпляр нецелесообразно или, когда закономерности в данных неизвестны. Этот процесс включает в себя выявление сходств или закономерностей в данных для объединения похожих точек данных в группы.

Ключевые шаги:

* Предварительная обработка данных: первым шагом является подготовка данных к анализу. Это включает в себя обработку пропущенных значений, масштабирование характеристик и удаление ненужной информации
* Выбор признаков для обучения
* Выбор алгоритма сегментации: для сегментации можно использовать несколько алгоритмов обучения без учителя, каждый из которых имеет свои сильные и слабые стороны. К распространенным методам относятся[кластеризация методом k-средних,](https://www.geeksforgeeks.org/k-means-clustering-introduction/) иерархическая кластеризация, DBSCAN (пространственная кластеризация приложений с шумом на основе плотности) и модели гауссовских смесей (GMM).
* Выбор алгоритма зависит от характера данных и желаемых характеристик кластеров.
* Выбор количества кластеров: некоторые алгоритмы, например, k-средних, требуют предварительного указания количества кластеров.
* Обучение модели: после выбора алгоритма и количества кластеров модель обучается на наборе данных.
* Оценка сегментации: хотя при обучении без учителя нет явных меток для оценки, существуют показатели, которые можно использовать для оценки качества сегментации. Для измерения сплоченности внутри кластеров и разделения между кластерами можно использовать показатели внутренней валидации, такие как оценка силуэта или индекс Дэвиса-Болдина.

На выходе получим:

* понимание аудитории, способность описать основные группы пользователей и их интересы
* выявление сегментов с максимальной монетизацией
* выбор маркетинговой стратегии
* налаживание эффективного взаимодействия с пользователями