**Проект: Обучение с учителем: качество модели**

**Описание проекта**

Интернет-магазин «В один клик» продаёт разные товары: для детей, для дома, мелкую бытовую технику, косметику и даже продукты. Отчёт магазина за прошлый период показал, что активность покупателей начала снижаться. Привлекать новых клиентов уже не так эффективно: о магазине и так знает большая часть целевой аудитории. Возможный выход — удерживать активность постоянных клиентов. Сделать это можно с помощью персонализированных предложений.

«В один клик» — современная компания, поэтому её руководство не хочет принимать решения просто так — только на основе анализа данных и бизнес-моделирования. У компании есть небольшой отдел цифровых технологий, и вам предстоит побыть в роли стажёра в этом отделе.

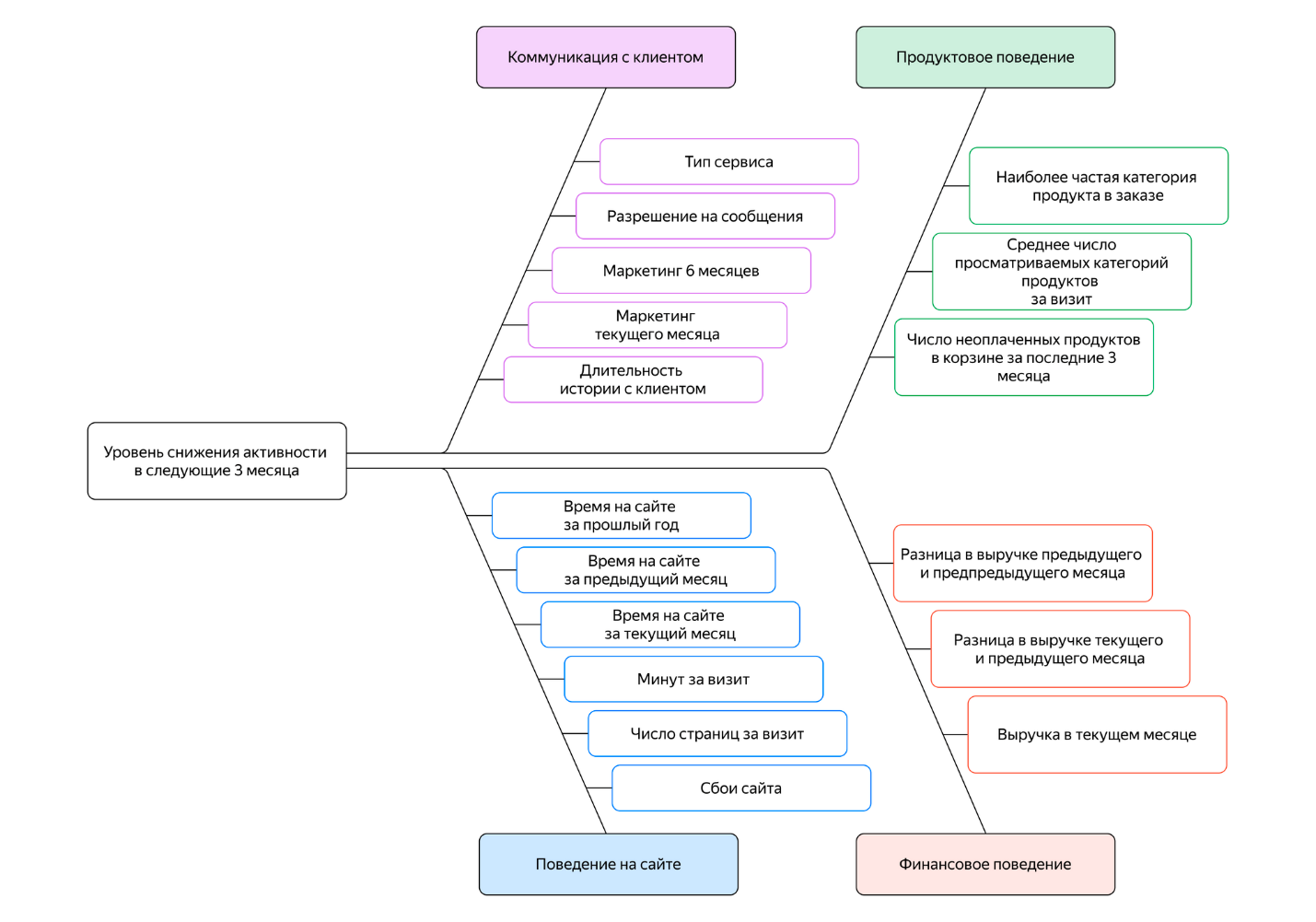
Итак, вашему отделу поручили разработать решение, которое позволит персонализировать предложения постоянным клиентам, чтобы увеличить их покупательскую активность.

**Как решать задачу**

Продумывать подход к решению этой задачи вам не придётся — руководитель отдела его уже сформировал:

1. Нужно промаркировать уровень финансовой активности постоянных покупателей. В компании принято выделять два уровня активности: «снизилась», если клиент стал покупать меньше товаров, и «прежний уровень».
2. Нужно собрать данные по клиентам по следующим группам:
   * Признаки, которые описывают коммуникацию сотрудников компании с клиентом.
   * Признаки, которые описывают продуктовое поведение покупателя. Например, какие товары покупает и как часто.
   * Признаки, которые описывают покупательское поведение клиента. Например, сколько тратил в магазине.
   * Признаки, которые описывают поведение покупателя на сайте. Например, как много страниц просматривает и сколько времени проводит на сайте.

Представим группы признаков (вместе с целевым) в виде диаграммы — такую визуализацию ещё называют диаграммой Исикавы.



1. Нужно построить модель, которая предскажет вероятность снижения покупательской активности клиента в следующие три месяца.
2. В исследование нужно включить дополнительные данные финансового департамента о прибыльности клиента: какой доход каждый покупатель приносил компании за последние три месяца.
3. Используя данные модели и данные о прибыльности клиентов, нужно выделить сегменты покупателей и разработать для них персонализированные предложения.

Руководство одобрило описание решения, и вам, как специалисту по DS, нужно его реализовать.

**Описание данных**

Данные для работы находятся в нескольких таблицах. Нажмите на название файла, чтобы его скачать.

[market\_file.csv](https://code.s3.yandex.net/datasets/market_file.csv)

Таблица, которая содержит данные о поведении покупателя на сайте, о коммуникациях с покупателем и его продуктовом поведении.

* id — номер покупателя в корпоративной базе данных.
* Покупательская активность — рассчитанный класс покупательской активности (целевой признак): «снизилась» или «прежний уровень».
* Тип сервиса — уровень сервиса, например «премиум» и «стандарт».
* Разрешить сообщать — информация о том, можно ли присылать покупателю дополнительные предложения о товаре. Согласие на это даёт покупатель.
* Маркет\_актив\_6\_мес — среднемесячное значение маркетинговых коммуникаций компании, которое приходилось на покупателя за последние 6 месяцев. Это значение показывает, какое число рассылок, звонков, показов рекламы и прочего приходилось на клиента.
* Маркет\_актив\_тек\_мес — количество маркетинговых коммуникаций в текущем месяце.
* Длительность — значение, которое показывает, сколько дней прошло с момента регистрации покупателя на сайте.
* Акционные\_покупки — среднемесячная доля покупок по акции от общего числа покупок за последние 6 месяцев.
* Популярная\_категория — самая популярная категория товаров у покупателя за последние 6 месяцев.
* Средний\_просмотр\_категорий\_за\_визит — показывает, сколько в среднем категорий покупатель просмотрел за визит в течение последнего месяца.
* Неоплаченные\_продукты\_штук\_квартал — общее число неоплаченных товаров в корзине за последние 3 месяца.
* Ошибка\_сервиса — число сбоев, которые коснулись покупателя во время посещения сайта.
* Страниц\_за\_визит — среднее количество страниц, которые просмотрел покупатель за один визит на сайт за последние 3 месяца.

[market\_money.csv](https://code.s3.yandex.net/datasets/market_money.csv)

Таблица с данными о выручке, которую получает магазин с покупателя, то есть сколько покупатель всего потратил за период взаимодействия с сайтом.

* id — номер покупателя в корпоративной базе данных.
* Период — название периода, во время которого зафиксирована выручка. Например, 'текущий\_месяц' или 'предыдущий\_месяц'.
* Выручка — сумма выручки за период.

[market\_time.csv](https://code.s3.yandex.net/datasets/market_time.csv)

Таблица с данными о времени (в минутах), которое покупатель провёл на сайте в течение периода.

* id — номер покупателя в корпоративной базе данных.
* Период — название периода, во время которого зафиксировано общее время.
* минут — значение времени, проведённого на сайте, в минутах.

[money.csv](https://code.s3.yandex.net/datasets/money.csv)

Таблица с данными о среднемесячной прибыли покупателя за последние 3 месяца: какую прибыль получает магазин от продаж каждому покупателю.

* id — номер покупателя в корпоративной базе данных.
* Прибыль — значение прибыли.

**Инструкция по выполнению проекта**

Решать задачу вам нужно в два этапа:

1. Разработайте модель, которая предскажет вероятность снижения покупательской активности.
2. Выделите сегмент покупателей, проанализируйте его и предложите, как увеличить его покупательскую активность. Используйте данные моделирования, данные о прибыли покупателей и исходные данные (если понадобятся). По желанию вы можете проанализировать больше одного сегмента. Ниже несколько примеров сегментов, но вы также можете выделить свой:
   * Группа клиентов с максимальной долей покупок по акции и высокой вероятностью снижения покупательской активности.
   * Группа клиентов, которые покупают только технику, то есть товары с длинным жизненным циклом.
   * Группа клиентов, которые покупают товары из категории «Товары для себя» (новая категория, которую можно выделить на основе текущих) или «Товары для детей».
   * Группа клиентов с высокой вероятностью снижения покупательской активности и наиболее высокой прибыльностью.

**Шаг 1. Загрузка данных**

1.1 Загрузите данные. Файлы market\_file.csv, market\_money.csv, market\_time.csv, money.csv доступны по ссылкам в описании данных или через пути к ним:

* /datasets/market\_file.csv
* /datasets/market\_money.csv
* /datasets/market\_time.csv
* /datasets/money.csv

Обратите внимание: в некоторых файлах разделитель значений — точка с запятой, а разделитель десятичных значений — запятая.

1.2 Проверьте, что данные в таблицах соответствуют описанию. Исследованием и объединением данных вы займётесь позже.

**Шаг 2. Предобработка данных**

Проведите необходимую предобработку данных. Названия столбцов можно оставить кириллическими.

**Шаг 3. Исследовательский анализ данных**

Проведите исследовательский анализ данных из каждой таблицы. Отберите клиентов с покупательской активностью не менее трёх месяцев, то есть таких, которые что-либо покупали в этот период. Оформите выводы по результатам шага.

**Шаг 4. Объединение таблиц**

4.1 Объедините таблицы market\_file.csv, market\_money.csv, market\_time.csv. Данные о прибыли из файла money.csv при моделировании вам не понадобятся.

4.2 Учитывайте, что данные о выручке и времени на сайте находятся в одном столбце для всех периодов. В итоговой таблице сделайте отдельный столбец для каждого периода.

**Шаг 5. Корреляционный анализ**

Проведите корреляционный анализ признаков в количественной шкале в итоговой таблице для моделирования. Сделайте выводы о мультиколлинеарности и при необходимости устраните её.

**Шаг 6. Использование пайплайнов**

Примените все изученные модели. Для этого используйте пайплайны.

6.1 Во время подготовки данных используйте ColumnTransformer. Количественные и категориальные признаки обработайте в пайплайне раздельно. Для кодирования категориальных признаков используйте как минимум два кодировщика, для масштабирования количественных — как минимум два скейлера.

Напоминаем, что для каждой модели можно подготовить данные с разным кодированием и масштабированием.

6.2 Обучите четыре модели: KNeighborsClassifier(), DecisionTreeClassifier(), LogisticRegression() и SVC(). Для каждой из них подберите как минимум один гиперпараметр. Выберите подходящую для задачи метрику, аргументируйте свой выбор. Используйте эту метрику при подборе гиперпараметров.

6.3 Выберите лучшую модель, используя заданную метрику. Для этого примените одну из стратегий:

* использовать пайплайны и инструменты подбора гиперпараметров для каждой модели отдельно, чтобы выбрать лучшую модель самостоятельно;
* использовать один общий пайплайн для всех моделей и инструмент подбора гиперпараметров, который вернёт вам лучшую модель.

**Шаг 7. Анализ важности признаков**

7.1 Оцените важность признаков для лучшей модели и постройте график важности с помощью метода SHAP.

7.2 Сделайте выводы о значимости признаков:

* какие признаки мало значимы для модели;
* какие признаки сильнее всего влияют на целевой признак;
* как можно использовать эти наблюдения при моделировании и принятии бизнес-решений.

**Шаг 8. Сегментация покупателей**

8.1 Выполните сегментацию покупателей. Используйте результаты моделирования и данные о прибыльности покупателей.

8.2 Выберите группу покупателей и предложите, как увеличить её покупательскую активность:

* Проведите графическое и аналитическое исследование группы покупателей.
* Сделайте предложения по работе с сегментом для увеличения покупательской активности.

8.3 Сделайте выводы о сегментах:

* какой сегмент вы взяли для дополнительного исследования,
* какие предложения вы сделали и почему.

**Шаг 9. Общий вывод**

Сделайте общий вывод:

* опишите задачу;
* опишите исходные данные и проведённую предобработку;
* напишите, что вы сделали для поиска лучшей модели;
* укажите лучшую модель;
* добавьте выводы и дополнительные предложения для выбранного сегмента покупателей.

**Оформление**

Выполните задание в Jupyter Notebook. Заполните программный код в ячейках типа code*,* текстовые пояснения — в ячейках типа markdown. Используйте форматирование и заголовки.

Обратите внимание: структуру исследования в Jupyter Notebook вам нужно создать самостоятельно.

**Как будут проверять мой проект**

Ваш проект будут оценивать по критериям. Прежде чем решать кейс, внимательно изучите их.

На что обращают внимание при проверке проектов:

* как вы описываете найденные в данных проблемы;
* какие методы построения графиков применяете для исследовательского анализа;
* какие методы предобработки данных применяете и как обосновываете принятое решение;
* какие ML-решения вы реализуете;
* как подбираете метрику и объясняете её выбор;
* как выбираете и объясняете лучшее ML-решение;
* как выполняете сегментацию покупателей, какой сегмент выбираете и какие решения по персонализации предлагаете бизнесу;
* соблюдаете ли структуру проекта и поддерживаете аккуратность кода;
* какие выводы делаете;
* оставляете ли комментарии к шагам.

Всё необходимое для того, чтобы выполнить проект, есть в темах, которые вы прошли.

Успехов!