# Постановка задачи.

Обработка и анализ данных:

Дан синтетический набор данных о продажах гипотетического интернет-магазина. Требуется выполнить следующие подзадачи:

* Провести предобработку данных (очистка, обработка пропусков, нормализация).
* Провести аналитический обзор данных с использованием SQL или Python (расчет основных метрик экономики интернет-магазина).
* Выявить гипотетические инсайты и предложить меры для увеличения эффективности продаж, основываясь на анализе.

Работа с ETL-процессами:

* Описать процесс загрузки данных из предоставленного датасета в Google BigQuery с использованием инструментов ETL.
* Требуется предоставить пример кода или описание логики работы ETL-пайплайна.

Визуализация данных:

* Создать интерактивный дашборд в Google Data Studio, отражающий ключевые метрики и инсайты, полученные в ходе анализа данных.
* Дашборд должен включать не менее 3-х видов визуализации (графики, таблицы, гистограммы и т.д.).

Дополнительное задание (для кандидатов, желающих продемонстрировать глубокие знания):

* Разработать простую модель машинного обучения на Python для прогнозирования спроса на наиболее популярные товары на основе исторических данных о продажах.
* Описать выбранный алгоритм, процесс обучения модели и оценить ее точность.

# Используемые технологии.

* Python.
* Google Data Studio.

# Обработка и анализ данных (код реализации Data\_science\_project.ipynb).

Введение

Код реализации приведен в Data\_science\_project.ipynb. Цель данного анализа заключается в оценке экономической эффективности интернет-магазина за 2022 год, идентификации ключевых факторов, влияющих на его продажи, и разработке стратегических рекомендаций для улучшения показателей. Анализ основан на данных о выручке, прибыли, конверсии и других важных экономических метриках.

Методология

Исходные данные включали информацию о продажах, выручке, прибыли и рекламных затратах интернет-магазина, разделенные по различным фирмам и рекламным кампаниям. Использовались библиотеки Python, включая pandas для обработки и агрегации данных и matplotlib для визуализации результатов. Основные шаги анализа включали предобработку данных, расчет ключевых метрик и выявление инсайтов.

Реализация

В процессе предобработки были удалены строки с отсутствующими датами сеансов, заполнены пропуски в числовых данных нулями, даты преобразованы в формат datetime, и удалены столбцы, не содержащие информативных данных (полностью нулевые столбцы). Эти шаги обеспечили корректность и полноту данных для анализа.

Анализ основных метрик

После подготовки данных, мы сосредоточимся на важных столбцах, чтобы рассчитать основные метрики экономики интернет-магазина. Получили следующие результаты:

* Общее количество показов: 554.744.457
* Общее количество кликов: 9.206.666
* Средний CTR (Click-Through Rate), %: 1.66%. Это означает, что примерно 1.66% показов приводят к клику.
* Средняя цена клика: 15.47. Это средняя стоимость одного клика по рекламе.
* Общие затраты: 228.126.116,96. Это сумма, потраченная на рекламу за период.
* Средний процент отказов: 19.08%. Процент сеансов, в которых пользователи покинули сайт сразу после перехода, составляет примерно 19.08%.
* Пример визуализации одной из метрик приведен на рисунке 1.

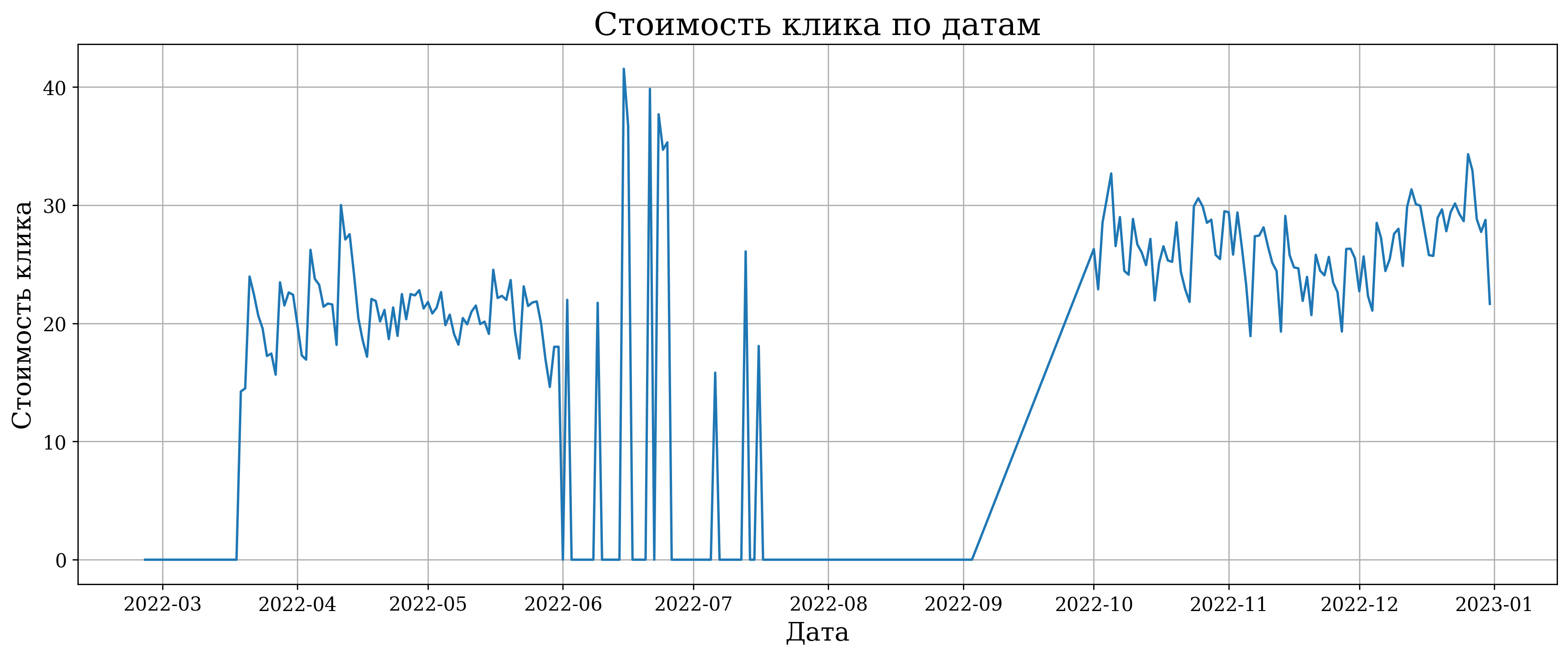


Рисунок 1. Пример графика относительно стоимости клика по датам

Основываясь на проведенном анализе, можно выявить несколько гипотетических инсайтов и предложить меры для увеличения эффективности продаж в интернет-магазине.

Гипотетические инсайты

1. Низкий CTR (1.66%) указывает на то, что рекламные объявления могут быть недостаточно привлекательными для целевой аудитории или не соответствовать контексту поиска.
2. Средняя цена клика 15.47 может быть сигналом к переоценке ставок в контекстной рекламе с учетом эффективности привлечения качественного трафика.
3. Высокие общие затраты на рекламу (228.126.116,96) при значительном проценте отказов (19.08%) говорят о возможной неэффективности некоторых рекламных кампаний.
4. Процент отказов 19.08% может свидетельствовать о несоответствии между ожиданиями пользователей и предложением на сайте, проблемах с удобством использования сайта или долгой загрузкой страниц.

Предложения по увеличению эффективности продаж

1. Оптимизация рекламных кампаний

* Пересмотреть таргетинг рекламных кампаний, чтобы точнее нацеливаться на целевую аудиторию.
* А/B тестирование различных рекламных сообщений и креативов для улучшения CTR.
* Использовать ремаркетинг для повторного привлечения пользователей, посетивших сайт, но не совершивших покупку.

1. Улучшение пользовательского опыта на сайте

* Анализ пользовательского поведения на сайте для выявления "узких мест" и оптимизации процесса покупки.
* Ускорение загрузки страниц сайта, особенно на мобильных устройствах.
* Улучшение дизайна сайта для повышения его удобства и привлекательности.

1. Персонализация предложения

* Разработать предложения и акции, ориентированные на интересы и предыдущие покупки пользователей.
* Использовать email-маркетинг для информирования о специальных предложениях и скидках.

1. Оптимизация продуктового ассортимента

* Анализ продаж для выявления наиболее и наименее популярных товаров.
* Корректировка ассортимента в соответствии с предпочтениями целевой аудитории.

1. Повышение конверсии через контент-маркетинг

* Создание полезного контента, который помогает потенциальным покупателям совершить осознанный выбор продукта.

Анализ метрик по разным источникам

Анализ выявил критические финансовые убытки на общую сумму 228.126.113,792, что говорит о значительных проблемах в экономической модели интернет-магазина. При этом, средние значения CPA и CR указывают на недостаточную эффективность рекламных инвестиций. Пример визуализации одной из метрик для всех источников отражен на рисунке 2.



Рисунок 2. Выручка по датам согласно разным источникам

Вывод и рекомендации

1. Немедленная оптимизация рекламных затрат

* Необходимо провести аудит всех рекламных кампаний, особенно тех, что приводят к высокому CPA и низкому CR, и скорректировать бюджеты в сторону более эффективных каналов.

1. Фокус на повышение конверсии

* Инвестировать в улучшение пользовательского опыта на сайте, A/B тестирование и оптимизацию воронки продаж для у величения конверсии. Важно анализировать поведение посетителей, выявлять и устранять препятствия на пути к покупке.

1. Улучшение управления ассортиментом

* Проанализировать ассортимент товаров на предмет их вклада в общую выручку и прибыль. Сфокусироваться на продвижении наиболее прибыльных товаров и категорий, а также рассмотреть возможность отказа от низкомаржинальных товаров.

1. Пересмотр стратегии ценообразования

* Исследовать возможности для оптимизации цен, возможно, через динамическое ценообразование, специальные предложения и скидки для увеличения среднего чека и привлечения новых клиентов.

1. Внедрение программ лояльности

* Разработать и внедрить программы лояльности для повышения удержания клиентов и стимулирования повторных покупок. Это может включать бонусы, накопительные скидки, персонализированные предложения.

1. Анализ и оптимизация логистики и операционных процессов

* Поиск путей снижения операционных затрат через оптимизацию логистических и операционных процессов, что может существенно повысить маржинальность.

# Работа с ETL-процессами (пример кода ETL\_process.ipynb).

Пример кода реализации приведен в ETL\_process.ipynb.

Подготовка

Необходимо убедиться в следующем:

1. Учетная запись Google Cloud и доступ к Google BigQuery.
2. Сервисный аккаунт с ключом в формате JSON, который предоставит вам доступ к Google Cloud API.
3. Установлены Python и необходимые библиотеки (pandas, google-cloud-bigquery, openpyxl).

1)Извлечение данных (Extract)

Чтение Excel файла

* Используем pandas для чтения данных из Excel файла. Необходимо убедиться, что указали правильный путь к файлу и листу, если в Excel файле их несколько.

2)Трансформация данных (Transform)

Преобразовываем данные, чтобы они соответствовали структуре и форматам предполагаемой целевой таблицы в BigQuery.

Данный этап может включать

* Преобразование типов данных (например, даты, числа).
* Очистка данных (удаление или исправление некорректных значений).
* Фильтрация ненужных строк или столбцов.
* Переименование столбцов для соответствия схеме целевой таблицы в BigQuery.

3)Загрузка данных (Load)

1. Настройка клиента BigQuery

* Необходимо использовать ключи имеющегося сервисного аккаунта для аутентификации и создайте клиент BigQuery.

1. Загрузка данных в BigQuery

* Затем используем метод load\_table\_from\_dataframe() клиента BigQuery для загрузки обработанных данных из pandas DataFrame в BigQuery.

Верификация

После выполнения этих шагов, можно проверить в предполагаемом проекте Google Cloud, что данные успешно загружены в указанную таблицу BigQuery. Можно выполнить SQL-запросы в BigQuery Console для проверки данных.

# Визуализация данных (Google data studio)

Прикрепляю ссылку на дашборд, доступ открыт по ссылке

<https://lookerstudio.google.com/reporting/a688ea97-201d-417f-a191-8ac892b59bb5>

# Дополнительное задание (реализация ML.ipynb).

Описание реализации

Код реализации приведен в ML.ipynb.

Для данной задачи было реализовано 3 различных метода:

1. Линейная регрессия.
2. Градиентный бустинг.
3. Случайный лес.

В качестве целевой переменной были выбраны "Показы", что позволяет оценить интерес к товару через количество раз, когда товар был показан пользователям. В результате оценки этих моделей с использованием коэффициента детерминации (R2) были получены следующие значения:

1. Линейная регрессия: R2 = 0.99273
2. Градиентный бустинг: R2 = 0.98972
3. Случайный лес: R2 = 0.97973

Эти результаты показывают, как хорошо модели объясняют вариативность целевой переменной "Показы". Высокие значения R2 для всех трех моделей указывают на то, что они могут достаточно точно предсказывать спрос на товары на основе предоставленных исторических данных.

Описание алгоритмов

1. Линейная регрессия предполагает линейную зависимость между входными переменными и целевой переменной. Это базовый алгоритм для прогнозирования, характеризующийся высокой интерпретируемостью и низкими вычислительными требованиями.
2. Градиентный бустинг строит модель, последовательно добавляя к ней слабые предсказательные модели (обычно деревья решений), так, чтобы они компенсировали недостатки уже добавленных моделей. Это позволяет достигать высокой точности, но требует тщательной настройки гиперпараметров.
3. Случайный лес является ансамблевым методом, который создает множество решающих деревьев на основе различных подвыборок обучающего набора данных и использует усреднение для улучшения точности прогнозирования и контроля переобучения.

Процесс обучения модели

Процесс обучения включал следующие шаги для каждого метода:

1. Предобработка данных

* подготовка и очистка данных, выбор релевантных признаков, обработка пропущенных значений.

1. Разделение данных

* разделение набора данных на обучающую и тестовую выборки.

1. Выбор и настройка модели

* настройка гиперпараметров выбранной модели.

1. Обучение модели

* обучение модели на обучающем наборе данных.

1. Оценка модели

* использование тестовой выборки для оценки точности модели с помощью метрики R2.

Оценка точности

Оценка точности моделей показала, что линейная регрессия обладает наивысшим значением R2, что делает ее предпочтительным выбором для данной задачи прогнозирования спроса на товары. Однако, градиентный бустинг и случайный лес также показали достаточно высокую точность, что делает их приемлемыми альтернативами в зависимости от специфики задачи, требований к интерпретируемости и вычислительных ресурсов (в случае большого набора данных или ограничений в вычислительных мощностях).