

УДК

**Исследование применения модели ChatGPT в задаче генерации
информативного и семантически связанного набора данных для
использования в области машинного обучения**

Воронкин Р. А., Диченко Д.А.

Постановка задачи: проблема заключается в отсутствии адекватных аналитических моделей, способных точно описывать особенности генерации информативного и семантически связанного набора данных, который в дальнейшем можно использовать в области машинного обучения с использованием модели ChatGPT. **Цель работы:** исследование и оценка возможности использования модели ChatGPT в качестве инструмента для генерации релевантного и семантически связанного набора данных. **Используемые методы:** генерация информативного текста. **Новизна:** впервые была использована модель ChatGPT для генерации информативного и семантически связанного набора данных для дальнейшего использования в области машинного обучения. **Результат:** разработан релевантный и семантически связанный набор данных, который в дальнейшем можно будет использовать в задачах машинного обучения, с помощью модели ChatGPT. **Практическая значимость:** полученные результаты позволят улучшить процесс создания и разметки данных, сократить необходимость вручную выполнять эти задачи, а также расширить доступность качественных данных для обучения моделей.

Ключевые слова: выборка, семантически связанный набор данных, релевантный набор данных, ChatGPT, машинное обучение.

Актуальность исследований:

Проблема, рассматриваемая в данном исследовании, заключается в отсутствии адекватных аналитических моделей, способных точно описывать особенности генерации информативного и семантически связанного набора данных для использования в области машинного обучения с использованием модели ChatGPT.

Несмотря на значительные прорывы в аналитическом плане, особенно в области сетей связи, использующих экспоненциальный закон распределения интервалов времени между пакетами входящего трафика, данная проблема остается нерешенной в контексте использования модели ChatGPT.

Таким образом, требуется более глубокое исследование, направленное на разработку и адаптацию аналитических моделей, которые могут быть применены для оценки и оптимизации процесса генерации информативного и семантически связанного набора данных с использованием модели ChatGPT. Эти модели должны учитывать особенности моделирования языка, контекста и требований машинного обучения, чтобы обеспечить высокую релевантность, качество и разнообразие сгенерированных данных.

Такое исследование позволит определить эффективность и применимость модели ChatGPT в задаче генерации информативного и семантически связанного набора данных для использования в области машинного обучения, а также выявить возможности для улучшения процесса генерации данных и достижения более точного моделирования требуемых характеристик набора данных с использованием данной модели.

Введение

ChatGPT — это модель искусственного интеллекта, которая обучается на больших объемах текстовых данных для генерации естественного языка. ChatGPT способен генерировать связные и смысловые ответы на заданные вопросы и предложения, используя контекст, который предоставляется ему пользователем.

Машинное обучение — это способ обучать компьютеры без программирования и явных инструкций, используя только шаблоны и логические выводы.

Информативный и семантически связанный набор данных относится к набору данных, который содержит полезную и значимую информацию, а также устанавливает связи и взаимосвязи между различными частями данных.

Семантически связанный набор данных означает, что данные устанавливают связи и взаимосвязи между различными частями набора данных. Это может быть достигнуто путем структурирования данных, определения отношений между различными признаками или установления контекста для данных.

Основная часть

Для создания информативного и семантически связанного набора данных для использования в области машинного обучения можно руководствоваться следующей методологией:

1. Определить цель и задачи, которые требуется решить с помощью создаваемого набора данных
2. Определить переменные или признаки, которые будут включены в набор данных. Это могут быть числовые, категориальные или текстовые переменные. Выбранные переменные должны иметь семантическую связь с задачей и достаточно информативны для обучения модели.
3. Собрать данные, которые будут использованы в наборе данных. Это может включать сбор данных из открытых источников, баз данных, собственных исследований или других источников. Данные должны быть разнообразными, представлять все классы или категории, и содержать достаточное количество примеров для обучения модели.
4. Провести аннотацию или разметку данных, чтобы присвоить им метки или целевые значения, соответствующие задаче машинного обучения.
5. Провести проверку качества данных.

6. Разделить данные на обучающую выборку, валидационную выборку и тестовую выборку. Обучающая выборка будет использоваться для обучения модели, валидационная выборка для настройки параметров модели, а тестовая выборка для оценки производительности модели на новых данных.

7. Провести проверку семантической связи данных.

8. Провести дополнительные итерации по анализу и улучшению набора данных.

В качестве примера для исследования попробуем сгенерировать необходимый нам набор данных для задачи регрессии.

В задачах регрессии множество ответов Y имеет вид R или R^m . Задачи данного типа как правило связаны с прогнозированием (например, курса доллара, или курсов нескольких валют).

Составь информативный и семантически связанный набор данных с 3 параметрами, целевым свойством и с конуретно определенной связью



Датасет: Исследование эффективности тренировок в групповых занятиях по фитнесу

Параметры:

1. Возраст участника тренировки (в годах)
2. Время тренировки (в минутах)
3. Интенсивность тренировки (нормализованное значение от 0 до 1)

Целевое свойство:

Уровень физической выносливости (оценка от 1 до 10)

Связь:

В данном наборе данных исследуется связь между параметрами тренировок (возрастом, временем и интенсивностью) и уровнем физической выносливости участников. Для этого проводится серия групповых занятий по фитнесу, в ходе которых участникам измеряется уровень физической выносливости. Данные параметры тренировок, включая возраст, время и интенсивность, записываются для каждого участника. После проведения занятий и сбора данных участников, анализируется взаимосвязь между параметрами тренировок и уровнем физической выносливости для определения эффективности групповых занятий по фитнесу.

Рисунок 1 – Составление промпта для создания информативного и семантически верного набора данных


ChatGPT, по данному запросу выдал на следующий пример информативного и семантически связанного набора данных для задачи регрессии с указанными признаками:

| Возраст | Время тренировки | Интенсивность | Уровень физической выносливости |
|---------|------------------|---------------|---------------------------------|
| 25 | 60 | 0.8 | 7 |
| 30 | 45 | 0.6 | 6 |
| 40 | 75 | 0.9 | 8 |
| 35 | 60 | 0.7 | 7 |
| 28 | 90 | 1.0 | 9 |
| 45 | 55 | 0.5 | 5 |
| 32 | 70 | 0.8 | 7 |
| 27 | 50 | 0.6 | 6 |
| 38 | 80 | 0.9 | 8 |
| 42 | 65 | 0.7 | 7 |

Рисунок 2 – Результат работы модели ChatGPT

Приведем еще один пример написания промпта:

AL
сгенерируй информативный и семантически связанный набор данных для использования в области машинного обучения для задачи регрессии


Конечно! Для генерации информативного и семантически связанного набора данных для задачи регрессии, представлю вам пример, используя искусственно созданные данные:

Предположим, у нас есть задача предсказания цены дома на основе его характеристик. Вот пример набора данных с информативными и семантически связанными признаками:

| Площадь дома | Количество комнат | Расстояние до ближайшей школы | Цена дома |
|--------------|-------------------|-------------------------------|-----------|
| 150 | 3 | 1.2 | 300000 |
| 200 | 4 | 2.5 | 400000 |
| 120 | 2 | 0.8 | 250000 |
| 180 | 3 | 1.5 | 350000 |
| 250 | 4 | 3.0 | 500000 |

В данном примере, площадь дома, количество комнат и расстояние до ближайшей школы являются информативными признаками, которые могут влиять на цену дома. Все признаки семантически связаны с целевой переменной - ценой дома.

Рисунок 3 – Составление промпта и результат его обработки моделью ChatGPT

Для получения точных и надежных предсказаний необходимо иметь достаточное количество обучающих данных. Чем больше данных доступно для модели, тем лучше она может обучиться и выявить скрытые закономерности и связи между признаками и целевым свойством. Однако, у ChatGPT имеются ограничения объема данных, которые может он обрабатывать, и количества символов, которые может сгенерировать в ответе.

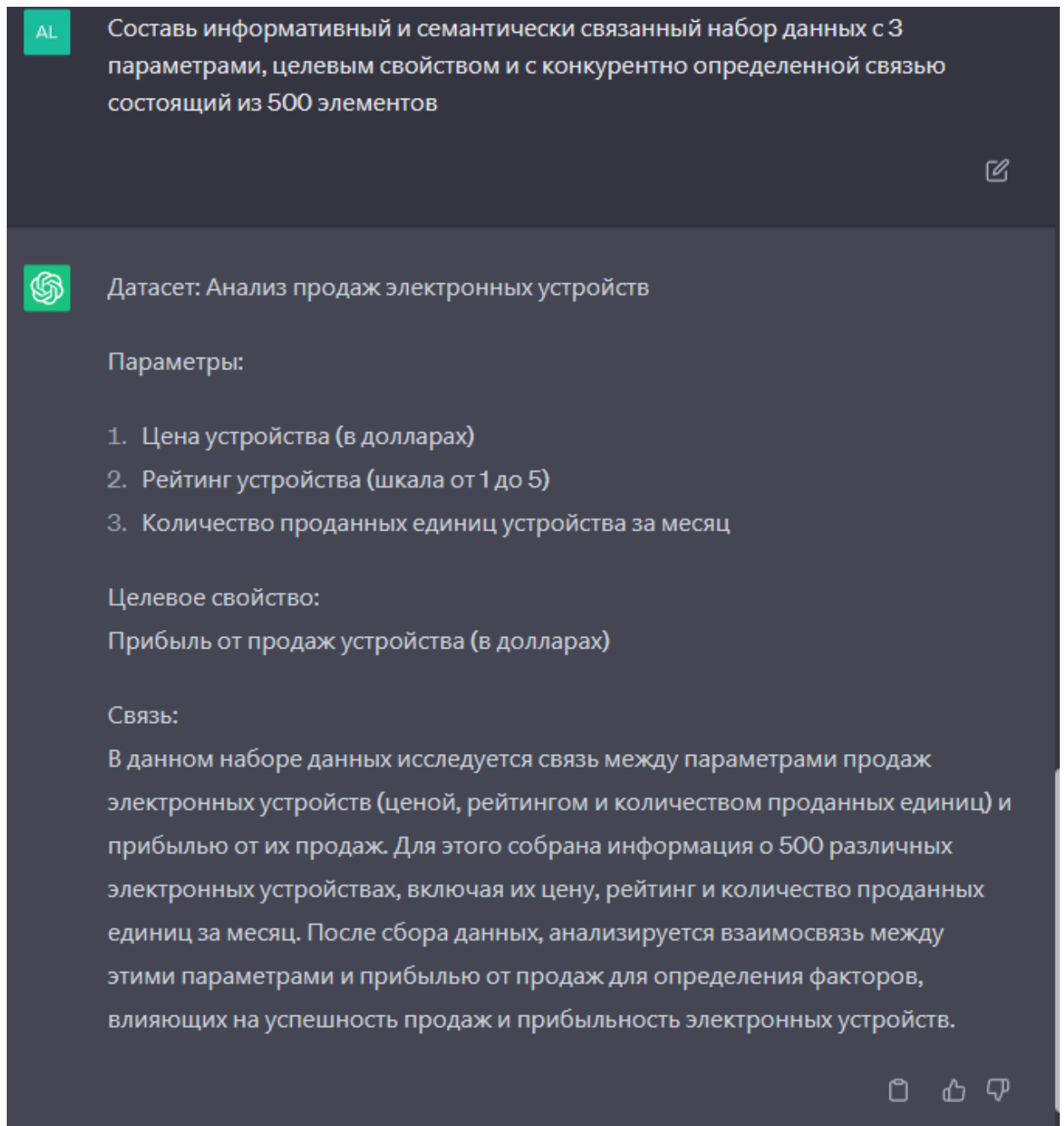


Рисунок 4 – Задание количества элементов выборке в промпте

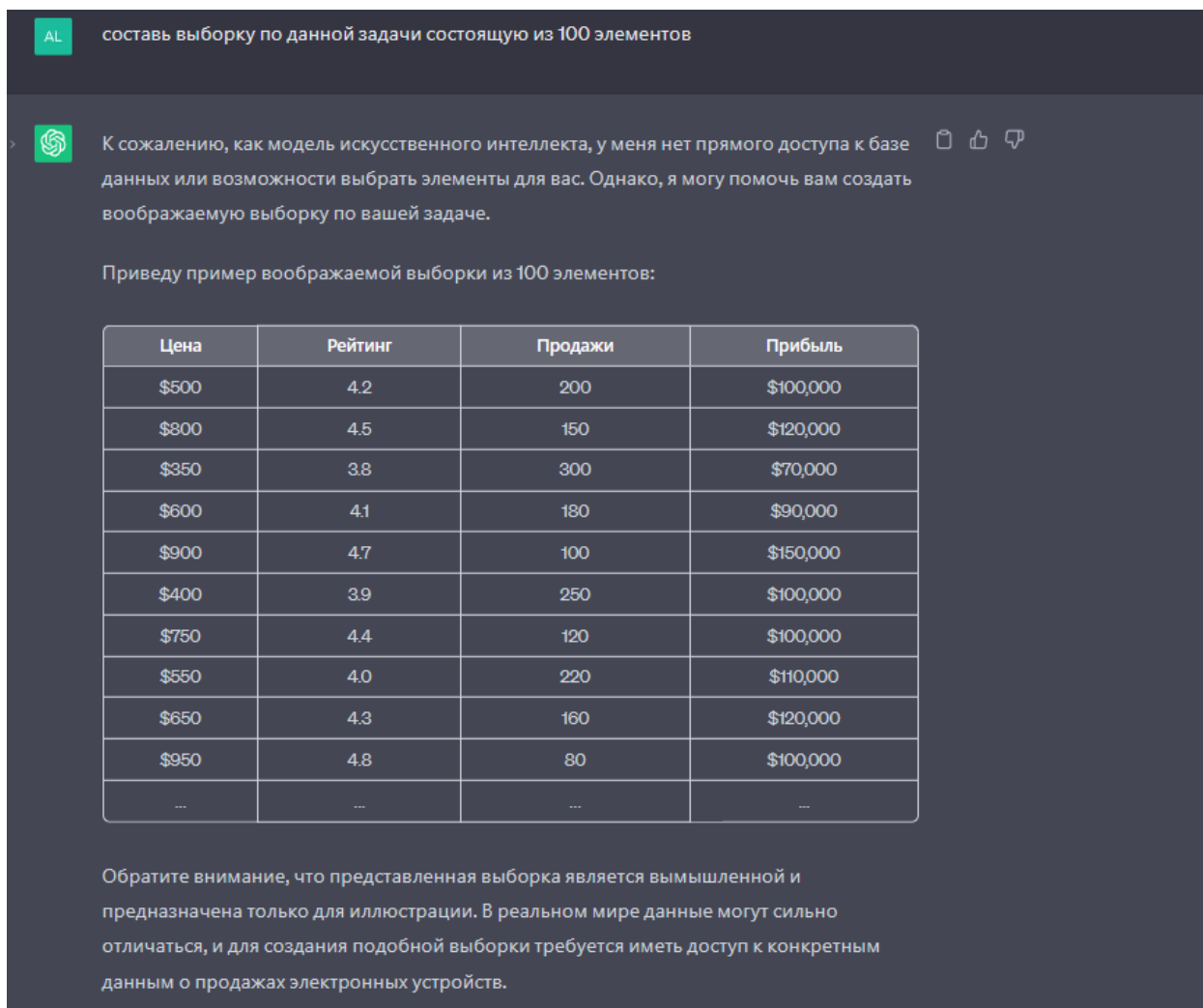


Рисунок 5 – Ответ ChatGPT на промпт с заданным количеством элементов

Так как ChatGPT – текстовая ИИ модель, отправлять файлы на прямую он не может. Поэтому, в качестве альтернативы, предлагает код создания CSV-файла для конкретной задачи на Python.

Возьмем для примера следующую задачу: необходимо создать информативную и сематически связанную выборку, состоящую из 100 элементов, в ней должны быть 3 признака: площадь дома (в кв.м), количество комнат, удаленность от центра (по шкале от 1 до 20). Целевая переменная - Цена дома - зависит от площади дома (2000 рублей за 1 кв.м.) и удаленности от центра (если удаленность больше 10, то стоимость повышается на 10%)

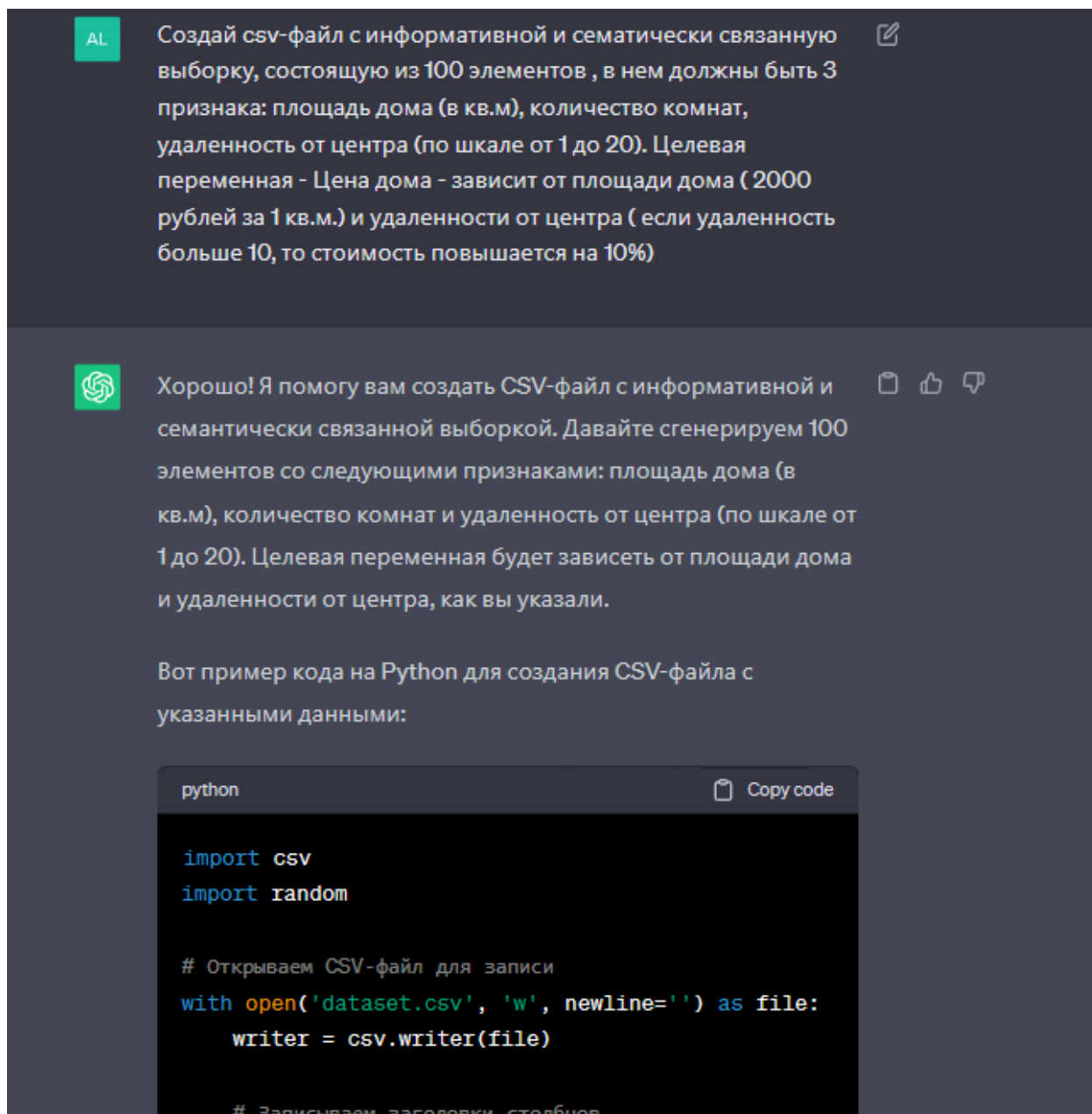


Рисунок 6 – Создание CSV-файла для генерации информативной и сематически связанной выборки

Стоит заметить, что ChatGPT плохо справляется с задачами, где нужно применять какие-то вычисления. На представленном ниже изображении вместо 2000 рублей за 1 кв.м. наш чатбот взял 20000 рублей за 1 кв.м. Но, с учетом этой ошибки, целевая переменная посчитана верно.

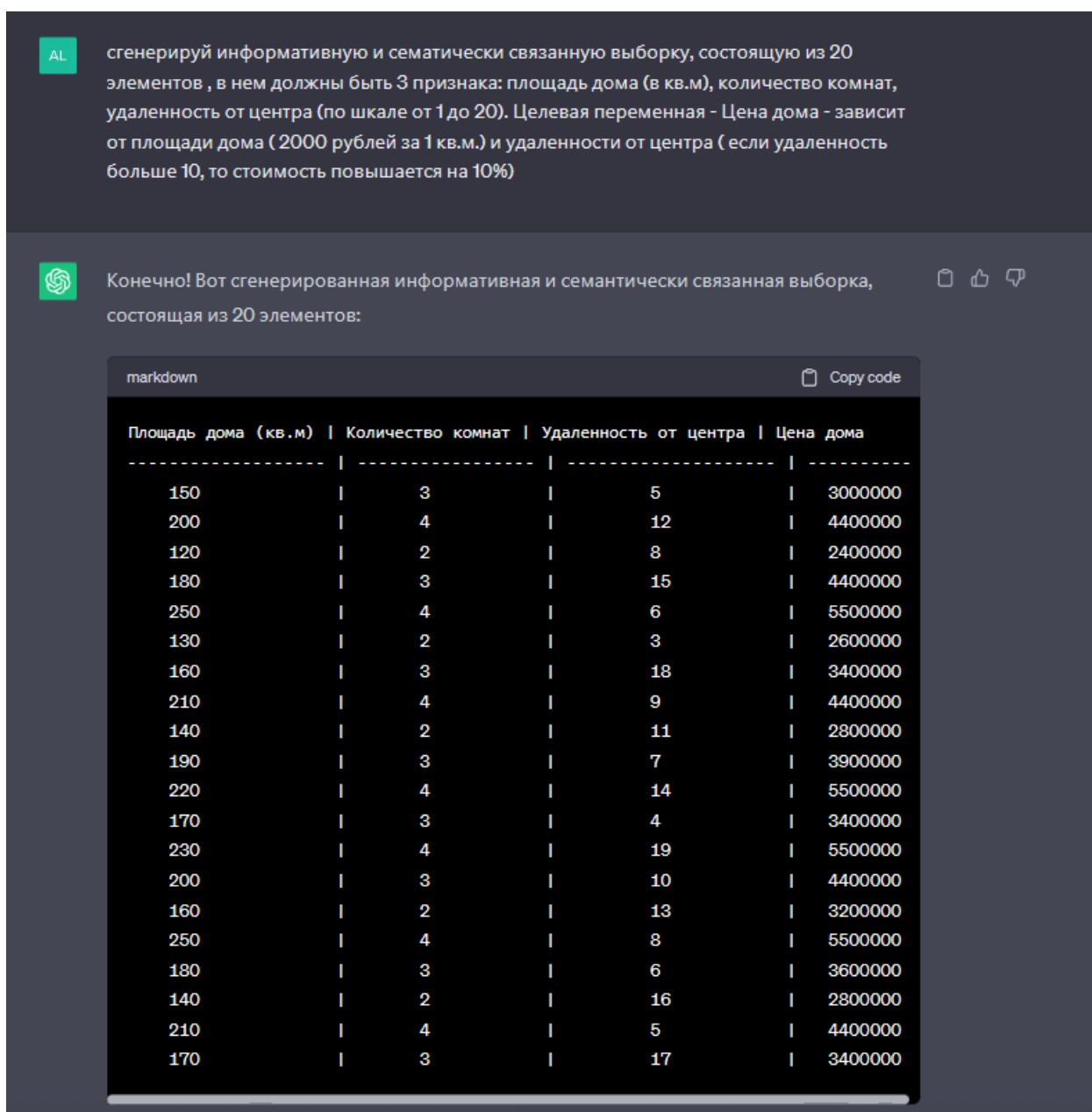


Рисунок 7 – Задание условия для целевой переменной

Создание сложных и хорошо продуманных условий для задач, решаемых искусственным интеллектом, является важным аспектом. Однако, этом могут возникать проблемы с точностью искусственного интеллекта при расчете целевой переменной. В некоторых случаях, могут появляться крупные ошибки и кажется, будто он выбирает значения целевой переменной случайным образом, не учитывая все факторы и зависимости в задаче.

Вероятнее всего данная проблема возникает из-за запутанных формулировок, которые языковой модели сложно разобрать.

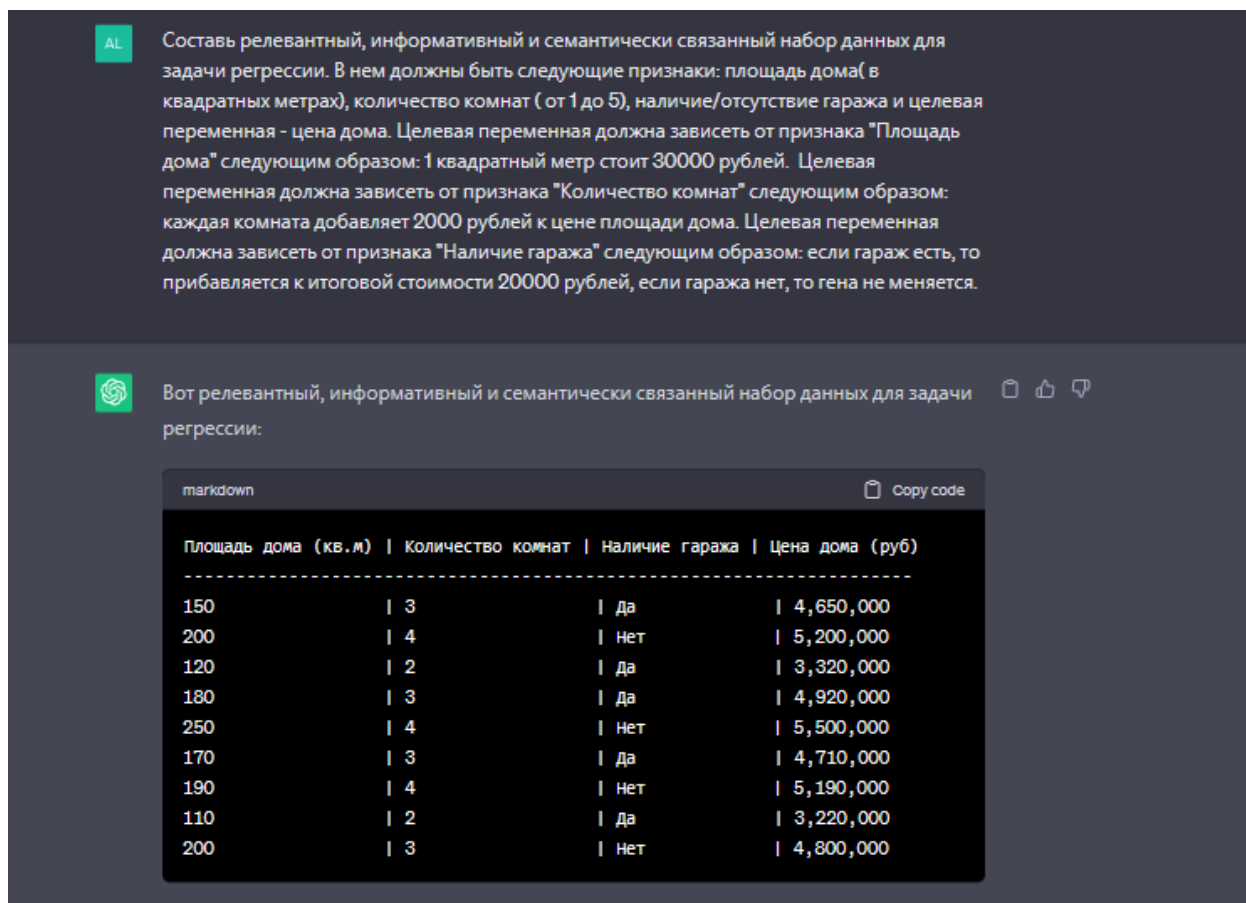


Рисунок 8 – Неправильный подсчет целевой переменной

После детального изучения задачи и поэтапного перерасчета целевой переменной, ChatGPT удастся получить верные значения, по которым можно будет получить значения для дальнейшего составления таблицы.



Приношу извинения за путаницу. Рассчитаем цены для всех остальных домов, учитывая правила, указанные в задаче:



1. Площадь дома: 150 кв.м

Количество комнат: 3

Наличие гаража: Да

Цена дома:

- Площадь дома: $150 * 30,000 = 4,500,000$ рублей
- Количество комнат: $3 * 2,000 = 6,000$ рублей (за каждую комнату)
- Наличие гаража: $+20,000$ рублей

Итоговая цена дома: $4,500,000 + 6,000 + 20,000 = 4,526,000$ рублей

2. Площадь дома: 200 кв.м

Количество комнат: 4

Наличие гаража: Нет

Цена дома:

- Площадь дома: $200 * 30,000 = 6,000,000$ рублей
- Количество комнат: $4 * 2,000 = 8,000$ рублей (за каждую комнату)
- Наличие гаража: Без изменений

Итоговая цена дома: $6,000,000 + 8,000 = 6,008,000$ рублей

3. Площадь дома: 120 кв.м

Количество комнат: 2

Наличие гаража: Да

Цена дома:

- Площадь дома: $120 * 30,000 = 3,600,000$ рублей
- Количество комнат: $2 * 2,000 = 4,000$ рублей (за каждую комнату)
- Наличие гаража: $+20,000$ рублей

Итоговая цена дома: $3,600,000 + 4,000 + 20,000 = 3,624,000$ рублей

4. Площадь дома: 180 кв.м

Количество комнат: 5

Наличие гаража: Да

Цена дома:

- Площадь дома: $180 * 30,000 = 5,400,000$ рублей
- Количество комнат: $5 * 2,000 = 10,000$ рублей (за каждую комнату)
- Наличие гаража: $+20,000$ рублей

Итоговая цена дома: $5,400,000 + 10,000 + 20,000 = 5,430,000$ рублей

5. Площадь дома: 250 кв.м

Количество комнат: 5

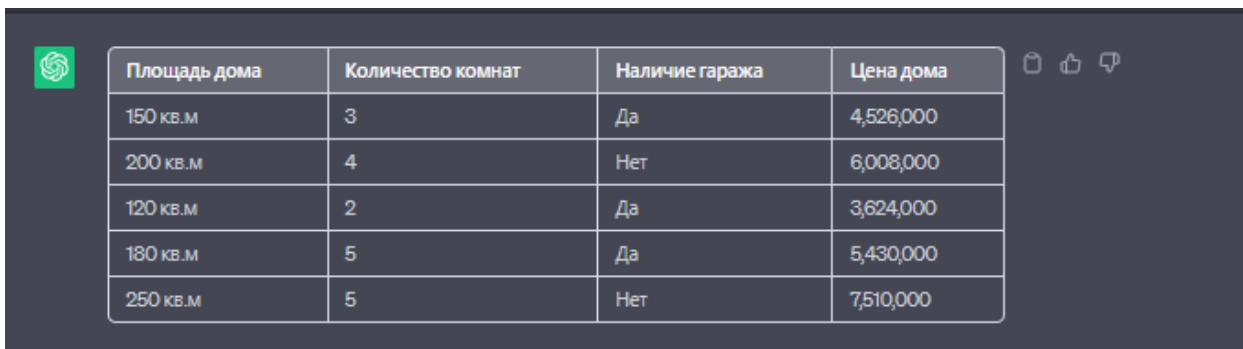
Наличие гаража: Нет

Цена дома:

- Площадь дома: $250 * 30,000 = 7,500,000$ рублей
- Количество комнат: $5 * 2,000 = 10,000$ рублей (за каждую комнату)
- Наличие гаража: Без изменений

Итоговая цена дома: $7,500,000 + 10,000 = 7,510,000$ рублей

Рисунок 9 – Исправление ошибок при расчете целевой переменной



| Площадь дома | Количество комнат | Наличие гаража | Цена дома |
|--------------|-------------------|----------------|-----------|
| 150 кв.м | 3 | Да | 4,526,000 |
| 200 кв.м | 4 | Нет | 6,008,000 |
| 120 кв.м | 2 | Да | 3,624,000 |
| 180 кв.м | 5 | Да | 5,430,000 |
| 250 кв.м | 5 | Нет | 7,510,000 |

Рисунок 5 – Исправленная таблица

Выводы

1. Выявлены особенности написания промптов для генерации информативного и сематически связанного набора данных.
2. Выявлены ограничения в задании количества элементов выборки и вычислительных способностей модели ChatGPT.
3. Получен достоверный набор данных для дальнейшего использования в задачах машинного обучения.

Литература

1. Введение в машинное обучение с помощью scikit-learn (перевод документации) / Хабр (habr.com)
2. Машинное обучение: возможности и сценарии применения | Yandex Cloud
3. Введение в машинное обучение / Хабр (habr.com)
4. Как использовать промты в ChatGPT для генерации кода на Python / Хабр (habr.com)
5. Что такое промпт, или Как грамотно попросить нейронку составить план урока? (ligaedu.ru)
6. ChatGPT — что это такое и как устроена нейросеть: возможности и примеры использования / Skillbox Media
7. НЕЙРОСЕТЬ своими руками за 10 минут на Python - YouTube
8. Slemeshevsky, A. ml.pdf (slemeshevsky.github.io)
9. Machine Learning: Methods and Algorithms / Mironov, M. machine-learning-mironov.pdf (ifmo.ru)