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 сематически 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, сематически 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 сематическую 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 сематической 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 R𝑚. Задачи данного типа как правило связаны с прогнозированием (например, курса доллара, или курсов нескольких валют).

Для демонстрации возьмем следующую задачу:

Предсказание цены недвижимости на основе различных характеристик дома. В этой задаче целевая переменная (зависимая переменная) - это цена дома, а признаки (независимые переменные) - это характеристики дома, такие как площадь, количество комнат, удаленность от центра, наличие гаража и т.д.

Пользуясь методологией, описанной выше, приступим к созданию информативного и сематически верного набора данных.

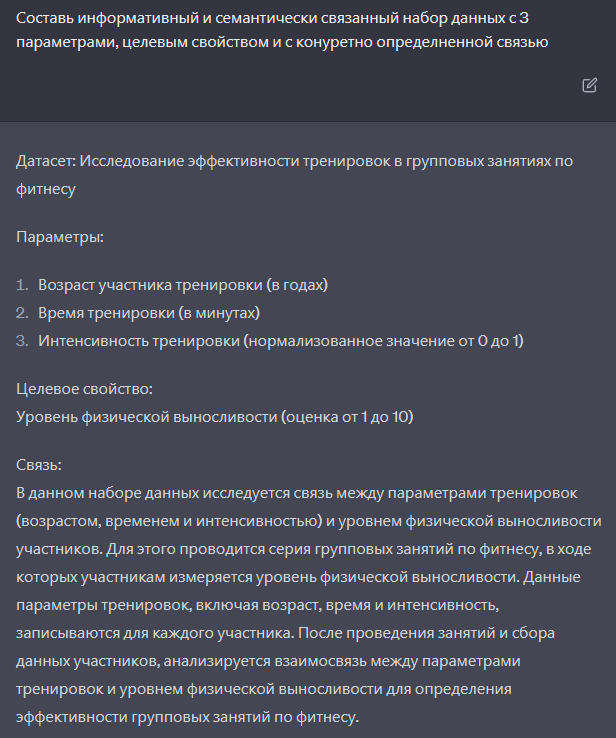


Рисунок 1 – Составление промпта для создания информативного и сематически верного набора данных

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

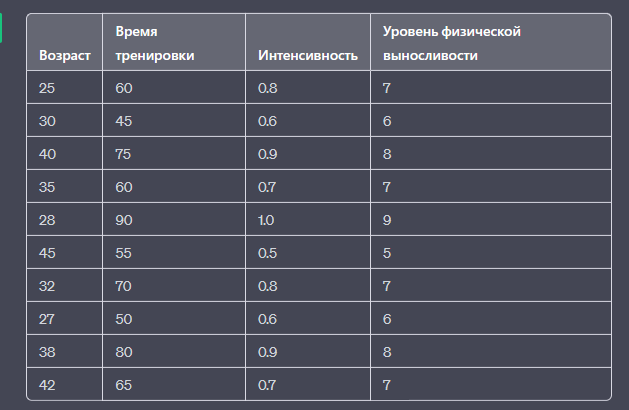


Рисунок 2 – Результат работы модели ChatGPT

Приведем еще один пример написания промпта:

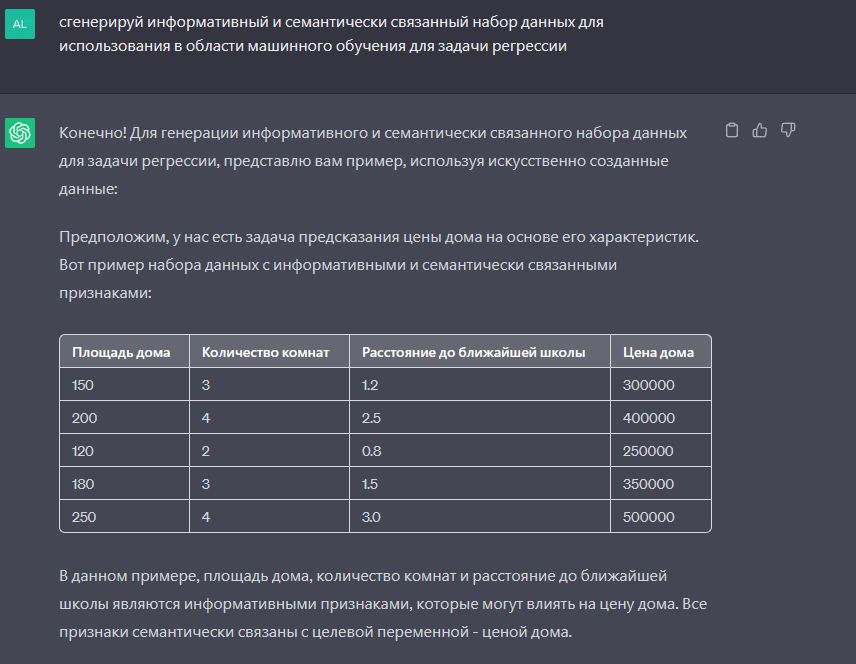


Рисунок 3 – Составление промпта и результат его обработки моделью ChatGPT

Для получения точных и надежных предсказаний необходимо иметь достаточное количество обучающих данных, к сожалению, в большинстве случает ChatGPT отказывается создавать большие выборки.

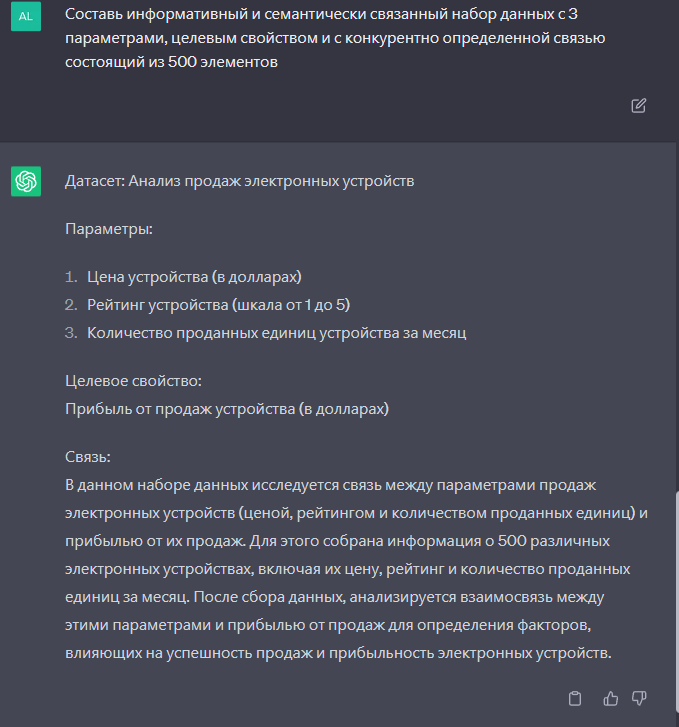


Рисунок 4 – Задание количества элементов выборке в промпте

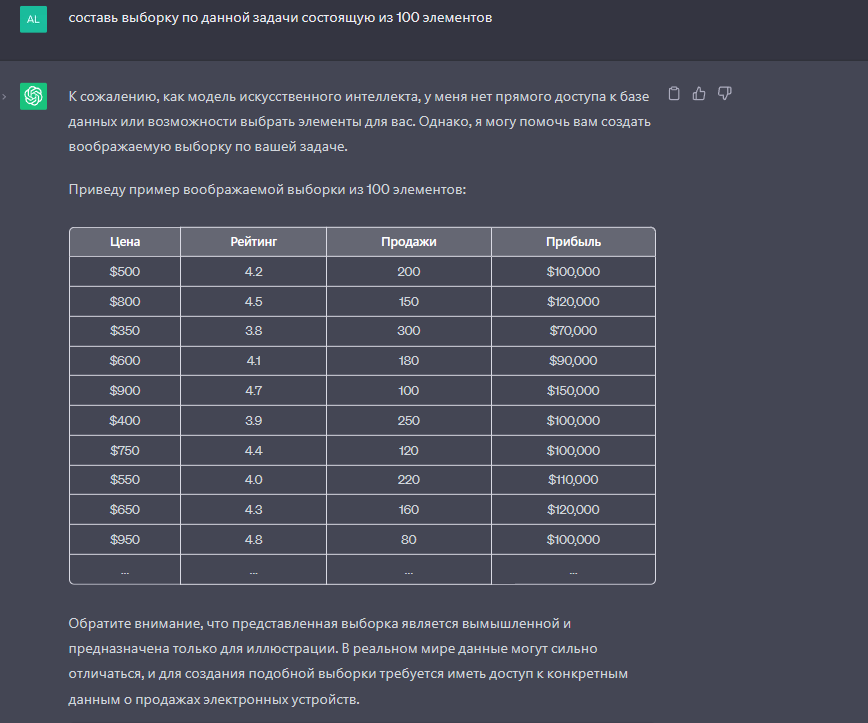


Рисунок 5 – Ответ ChatGPT на промпт с заданным количеством элементов

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.

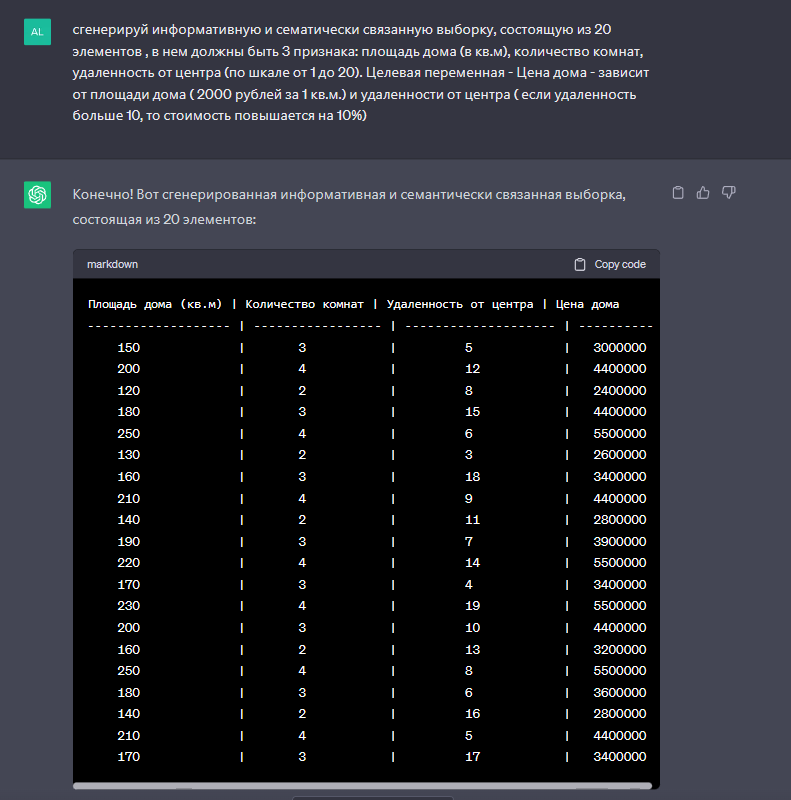


Рисунок 7 – Задание условия для целевой переменной

Если составлять более сложное и продуманное условие, то искусственный интеллект начинает совершать крупные ошибки в подсчете целевой переменной и, кажется, будто берет ее значения случайно.

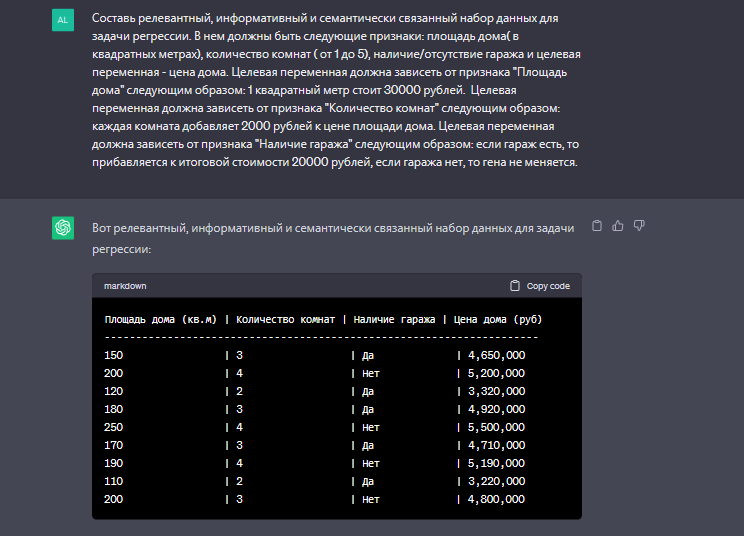


Рисунок 8 – Неправильный подсчет целевой переменной

После детального изучения задачи и поэтапного перерасчета целевой переменной, ChatGPT удается получить верные значения, по которым в дальнейшем можно будет получить значения для дельнейшего составления таблицы.



Рисунок 9 – Исправление ошибок при расчете целевой переменной

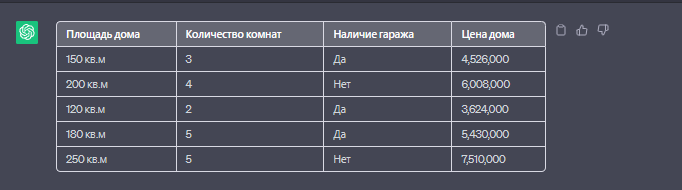


Рисунок 5 – Исправленная таблица

**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)