ЛАБОРАТОРНАЯ РАБОТА 9. ИСПОЛЬЗОВАНИЕ РАЗРАБОТАННОГО ПАЙПЛАЙНА ДЛЯ МНОГОМЕРНОЙ РЕГРЕССИИ

**Исследование набора данных**

**Цели и задачи**

Цель лабораторной работы: научиться применять разработанный пайплайн для тиражирования кода с целью решения широкого круга задач машинного обучения.

Основные задачи:

* получение навыков рефакторинга кода в проектах машинного обучения;
* получение навыков определения ключевых признаков в задачах машинного обучения;
* получение навыков реализации ключевых стратегий оптимизации моделей регрессии.

**Индивидуальное задание**

**Описание набора данных**

Данный набор данных является коллекцией пользовательских данных из социальных сетей. Он содержит информацию о отдельных пользователях, их паттернах использования и характеристиках.

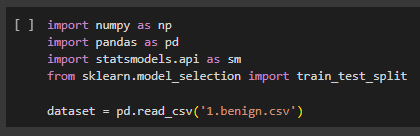
Цель набора данных:

Целью этого набора данных является анализ поведения пользователей в социальных сетях, понимание их поведения, использования и выявление трендов. Этот набор данных может быть использован для построения моделей, предсказывающих вовлеченность пользователей, идентификации пользователей в сети и оптимизации маркетинга в социальных сетях.

**Индивидуальное задание**

Постройте модель многомерной регрессии с использованием стратегии backward elimination.

1 Подключение библиотек и загрузка датасета:



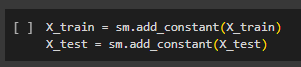
2 Определение матрицы признаков и целевой переменной



3 Разделение данных на обучающую и тестовую выборки



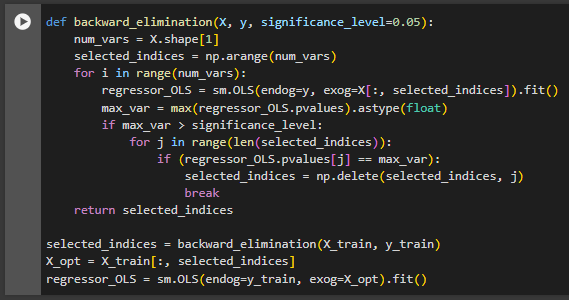
4 Добавление константного столбца для OLS



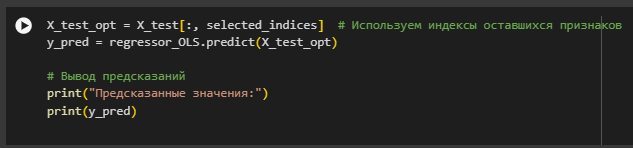
5 Построение модели OLS



6 Применение метода backward elimination

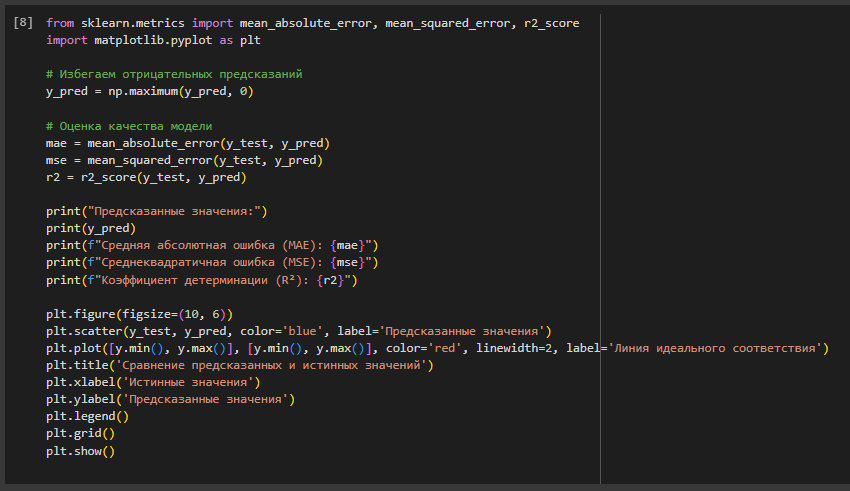


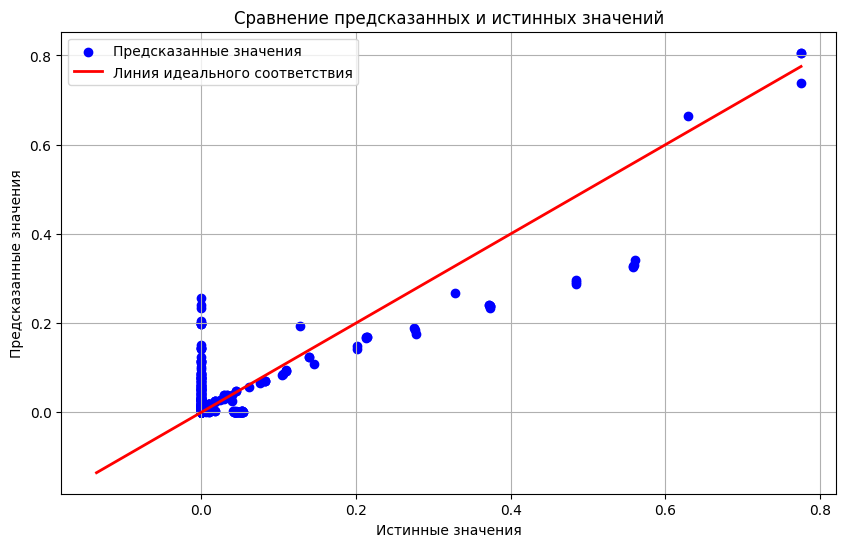
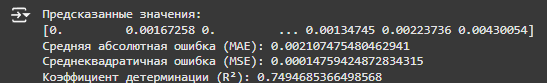
7 Предсказание на тестовой выборке и вывод предсказанных значений





8 Оценка качества модели и визуализация предсказанных и истинных значений





**Контрольные вопросы**

1. Фиктивный признак (константа) добавляется для учета свободного члена в уравнении регрессии. Это позволяет модели корректно интерпретировать влияние всех признаков.
2. Фиктивная переменная (dummy variable) используется для представления категориальных признаков в числовом формате. При перекодировке категориального признака создаются фиктивные переменные, но одна из них удаляется, чтобы избежать мультиколлинеарности (избыточной корреляции) между переменными.
3. Удаляемый признак выбирается на основе p-значения: признак с наибольшим p-значением, превышающим заданный уровень значимости (обычно 0.05), удаляется из модели.
4. Алгоритм all-in regression включает все доступные признаки в модель с самого начала и затем оценивает их значимость, чтобы определить, какие из них следует оставить, а какие удалить.
5. Алгоритм forward selection regression начинает с пустой модели и последовательно добавляет признаки, основываясь на их значимости, пока не будет достигнут оптимальный набор признаков.
6. Алгоритм Bidirectional Elimination сочетает элементы forward selection и backward elimination, позволяя одновременно добавлять и удалять признаки, оптимизируя модель.
7. Удаление признаков можно реализовать автоматически, используя цикл, который на каждой итерации проверяет p-значения всех признаков и удаляет тот, который имеет наибольшее p-значение, превышающее заданный уровень значимости, до тех пор, пока все оставшиеся признаки будут значимыми.