Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

|  |  |
| --- | --- |
| Институт | математики и компьютерных наук |
| Кафедра | компьютерной безопасности |

ОТЧЕТ

ЛАБОРАТОРНАЯ РАБОТА 7.

ИСПОЛЬЗОВАНИЕ РАЗРАБОТАННОГО ПАЙПЛАЙНА ДЛЯ МНОГОМЕРНОЙ РЕГРЕССИИ

Выполнил: Окунев Николай Александрович,

студент 2 курса

группы КМБ-с-о-23-1

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025 г.

Проверено с оценкой:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025 г.

Ставрополь, 2025

1. **Цели и задачи**

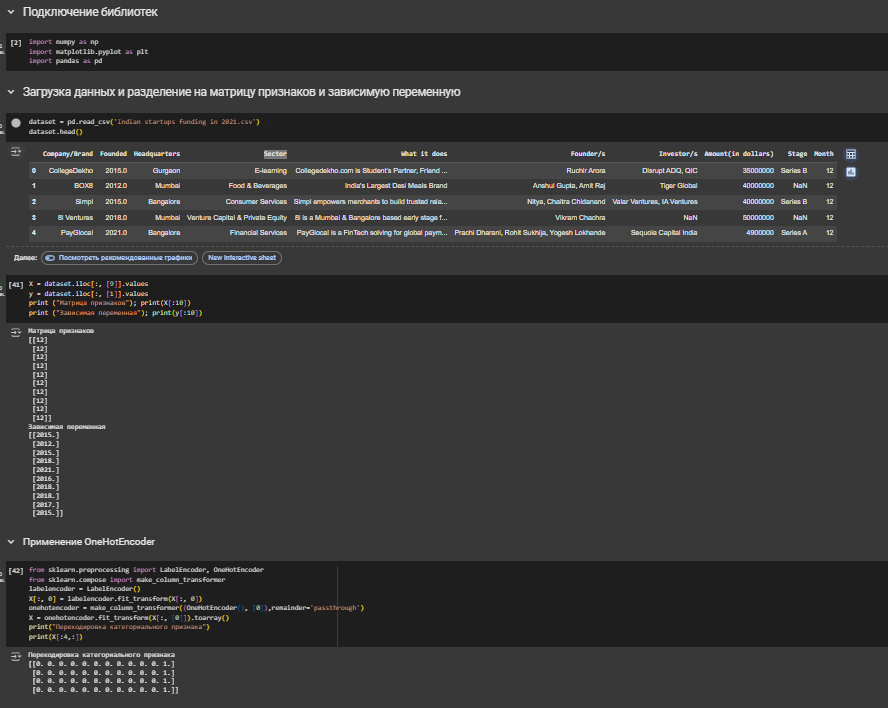
Цель лабораторной работы: научиться применять разработанный пайплайн для тиражирования кода с целью решения широкого круга задач машинного обучения. Основные задачи: – получение навыков рефакторинга кода в проектах машинного обучения; – получение навыков определения ключевых признаков в задачах машинного обучения; – получение навыков реализации ключевых стратегий оптимизации моделей регрессии.

1. **Теоретическое обоснование**

При решении задач многомерной регресси исследователю необходимо решить ряд подзадач:

1. Определить коррелированность признаков.
2. Определить, какие признаки существенны при построении модели регрессии. Проблема определения значимых признаков мвязана с проблемой снижения размерности. Важное значение при многомерной регресси приобретает обработка категориальных признаов. Часто необходимо заменить категориальный признак на набор фиктивных переменных.К проблеме выбора значимых переменных существует несколько стратегий (фактически это методы построения модели многомерной регрессии):
3. All-in. В данном подходе производится включение веех признаков в модель.
4. Backward Elimination. В подходе предполагается обучение модели с учетом всех признаков и удаление признаков по одному на основе их значимости до достижения ситуации, когда останутся только значимые признаки.
5. Forward Selection. Подход предполагает начальное тестирование модели с одним признаком (тестируется каждый признак). Затем добавляются по одному наиболее значимые признаки.
6. Bidirectional Elimination. Подход совмещает стратегии 2 и 3. 5. Score Comparison.
7. **Методика и порядок выполнения работы**

Перед выполнением индивидуального задания рекомендуется выполнить все пункты учебной задачи.

* 1. Учебная задача Задание. На основе разработанного пайплайна для линейной одномерной регрессии разработать многомерную модель регрессии. Решение. Для разработки модели необходимо реализовать следующий код: 

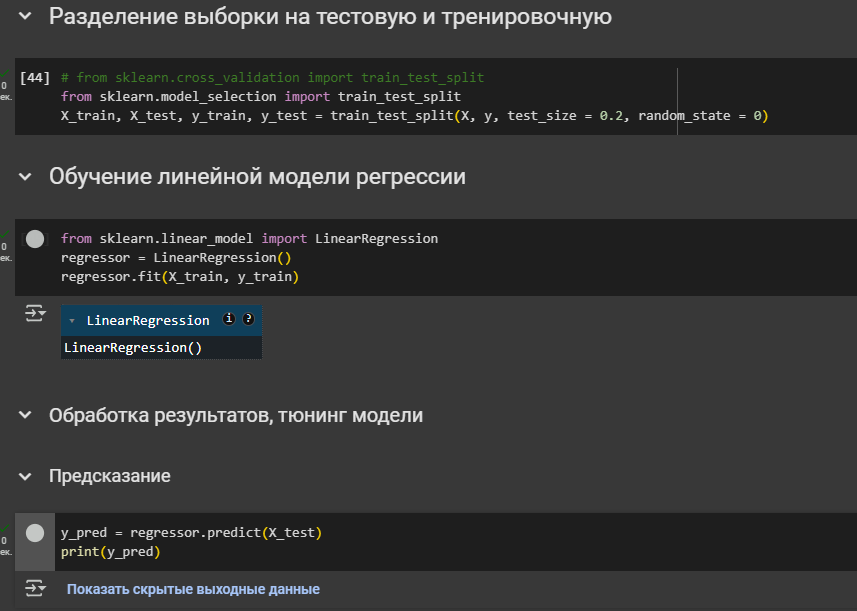


Рисунок 7.1 – Код Python для построения модели многомерной регрессии

На данный момент произведено обучение модели на всем наборе признаков. Для оптимизации модели реализуем стратегию Back Elimination (рисунк 7.2).

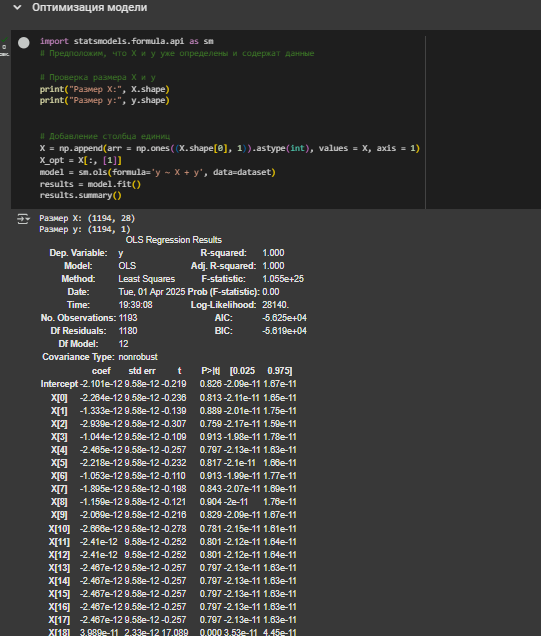


Рисунок 7.2 – Оптимизация модели многомерной регрессии

**Контрольные вопросы**

1)Фиктивный признак с 1.0 в многомерной линейной регрессии:

Этот фиктивный признак (столбец из единиц) нужен для учета свободного члена (intercept) в уравнении линейной регрессии. Уравнение линейной регрессии выглядит так: y = b0 + b1\*x1 + b2\*x2 + ... + bn\*xn где:

y — зависимая переменная

x1, x2, ..., xn — независимые переменные

b0 — свободный член (intercept)

b1, b2, ..., bn — коэффициенты регрессии

Без фиктивного признака модель будет проходить через начало координат (0, 0), что часто не соответствует реальным данным. Фиктивный признак позволяет модели смещаться относительно начала координат и находить наилучшую линию (или гиперплоскость), которая соответствует данным.

2)Фиктивная переменная и удаление одной фиктивной переменной:

Фиктивная переменная (dummy variable) — это числовая переменная, которая используется для представления категориального признака. Обычно для каждой категории создается отдельная фиктивная переменная, которая принимает значение 1, если объект принадлежит к этой категории, и 0 в противном случае. При перекодировке категориального признака с k категориями создается k фиктивных переменных. Однако, чтобы избежать мультиколлинеарности (сильной корреляции между независимыми переменными), одну из фиктивных переменных обычно удаляют. Мультиколлинеарность может привести к нестабильности оценок коэффициентов регрессии и затруднить интерпретацию результатов. Удаление одной фиктивной переменной не влияет на качество модели, но позволяет избежать мультиколлинеарности.

3)Критерий выбора удаляемого признака в Back Elimination:

В алгоритме Back Elimination удаляемый признак выбирается на основе p-значения (p-value). P-значение показывает вероятность получить наблюдаемые результаты (или более экстремальные), если нулевая гипотеза верна. В контексте регрессии нулевая гипотеза состоит в том, что коэффициент регрессии для данного признака равен 0 (то есть, признак не влияет на зависимую переменную). Чем выше p-значение, тем больше вероятность того, что нулевая гипотеза верна, и тем менее значим данный признак. В Back Elimination удаляется признак с наибольшим p-значением, которое превышает заданный уровень значимости (обычно 0.05).

4)All-in Regression:

All-in regression (или "force entry") — это метод построения модели регрессии, при котором все доступные признаки включаются в модель одновременно, без какого-либо предварительного отбора. Этот метод подходит, если у тебя есть веские основания полагать, что все признаки важны, или если у тебя мало признаков.

5)Forward Selection Regression:

Forward selection — это итеративный метод построения модели регрессии, который начинается с пустой модели (без признаков) и постепенно добавляет признаки, которые наиболее сильно улучшают качество модели. На каждом шаге добавляется признак, который имеет наименьшее p-значение (или наибольшую корреляцию с зависимой переменной, если p-значения недоступны). Процесс продолжается до тех пор, пока добавление новых признаков не перестает значительно улучшать модель.

6)Bidirectional Elimination:

Bidirectional elimination — это комбинация Forward selection и Backward elimination. Метод начинается с пустой модели (как Forward selection) и добавляет признаки, которые наиболее сильно улучшают качество модели. Однако, после каждого добавления признака метод также проверяет, нет ли каких-либо признаков, которые стали незначимыми и которые можно удалить (как Backward elimination). Этот процесс продолжается до тех пор, пока добавление новых признаков не перестает значительно улучшать модель, и удаление существующих признаков не перестает улучшать модель.

**Вывод:**научились применять разработанный пайплайн для тиражирования кода с целью решения широкого круга задач машинного обучения.