**Московский авиационный институт**

**(Национальный исследовательский университет)**

Факультет прикладной математики и физики

Кафедра вычислительной математики и программирования

**Курсовая работа**

по курсу «Методы, средства и технологии мультимедиа»

Студент: Катермин В. С.

Группа: М8О-408Б-18

Преподаватель: Б. В. Вишняков.

Оценка:

Москва, 2021

**Постановка задачи:**  Выбрать задачу (классификация или регрессия), датасет и метрику качества. Выбранные данные необходимо визуализировать и проанализировать. После этого выполнить препроцессинг. Затем реализовать алгоритм линейной регрессии, проверить качество обучения, сравнить с моделью из sklearn.

**Вариант:** Линейная регрессия. Будем предсказывать оценку университета на основании остальных признаков.

**Входные данные**

Датасет представляет из себя информацию об лучших университетах со всего мира. Содержит следующие признаки:

* Место в мире
* Название
* Страна
* Место в стране
* Качество образования
* Трудоустройство выпускников
* Качество преподавательского состава
* Публикации
* Влияние
* Цитаты
* Широкое влияние
* Патенты
* Оценка
* Год оценки

**Описание**

**Линейная регрессия** — модель зависимости переменной x от одной или нескольких других переменных (факторов, регрессоров, независимых переменных) с линейной функцией зависимости.

Линейная регрессия относится к задаче определения «линии наилучшего соответствия» через набор точек данных и стала простым предшественником нелинейных методов, которые используют для обучения нейронных сетей. В этой статье покажем вам примеры линейной регрессии.

**Градиентный спуск** — метод численной оптимизации, который может быть использован во многих алгоритмах, где требуется найти экстремум функции — нейронные сети, SVM, k-средних, регрессии. Однако проще его воспринять в чистом виде (и проще модифицировать).

**Регуляризация** — метод добавления некоторых дополнительных ограничений к условию с целью решить некорректно поставленную задачу или предотвратить переобучение. Чаще всего эта информация имеет вид штрафа за сложность модели.

**Регуляризация L2** – это тип линейной регрессии, который позволяет регуляризовать модель, основан на выборе как можно меньших значений веса. Другими словами, регуляризация ограничивает модель, уменьшая влияние входов на выход, тем самым модель становится регуляризованной и избегает переоснащения этими ограничениями.

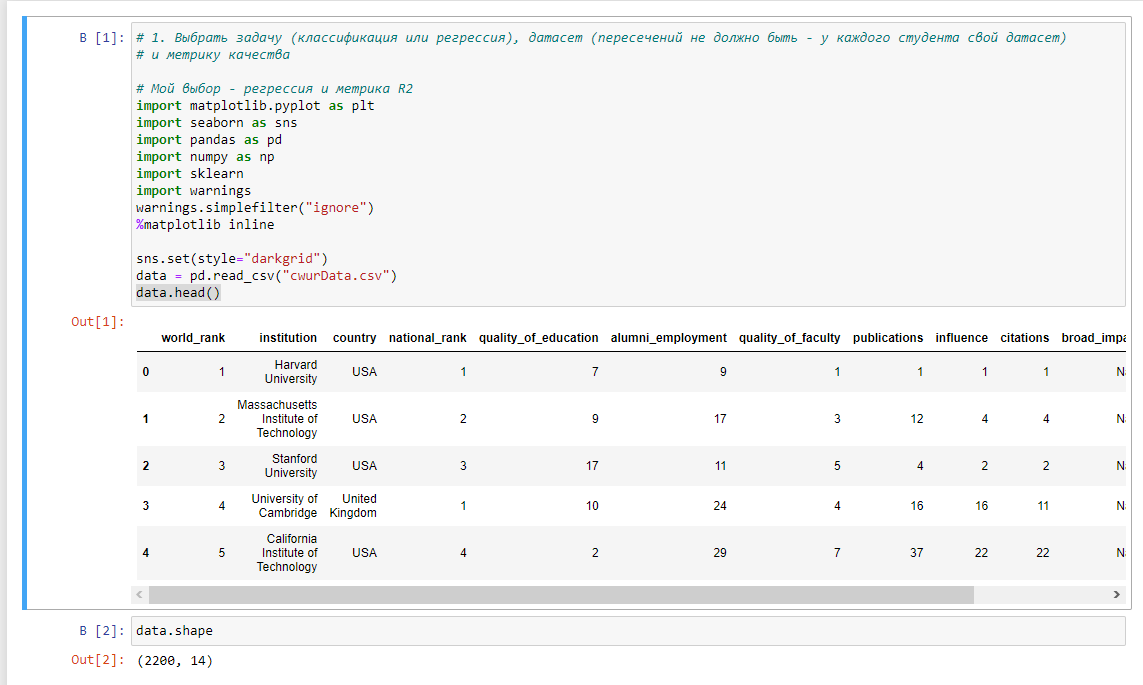
**Регуляризация L1** – оператор наименьшей абсолютной усадки и выбора (ЛАССО), при этом методе веса некоторых значений принимаются равными модулю числа, предполагая, что они не влияют на результат.

Разница между L1 и L2:

При L2-регуляризации дополнительный член является квадратичной функцией, при L1-регуляризации – модулем. Что здесь действительно важно – производная функции. Производная, конечно, является ключевой, поскольку градиентный спуск в основном движется в направлении производной.При квадратичном члене чем ближе вы находитесь к нулю, тем меньшей становится ваша производная, пока также не приблизится к нулю. Поэтому при L2-регуляризации когда ваша величина w уже мала, дальнейший градиентный спуск уже её сильно не изменит. В случае модуля производная является константой с абсолютной величиной, равной единице. Формально в нуле она не определена, но мы считаем её также равной нулю.Поэтому при L1-регуляризации градиентный спуск будет стремиться к нулю с постоянной скоростью, а достигнув его, там и останется. Вследствие этого L2-регуляризация способствует малой величине весовых коэффициентов, а L1-регуляризация способствует их равенству нулю, тем самым провоцируя разрежённость.

**Ход работы**

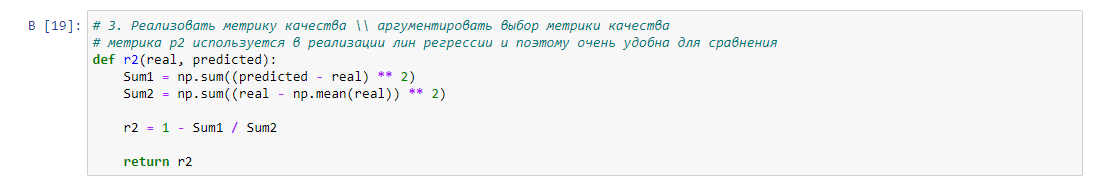
Загружаем датасет и изучаем его данные:

 Производим One Hot Encoding на категориальные признаках, однако у нас получится очень много “шума” – данных со значениями коэффициента корреляции меньше 0,1 по модулю. Поэтому все шумы мы переведём в отдельный столбец other, а сами шумовые столбцы — удаляем.

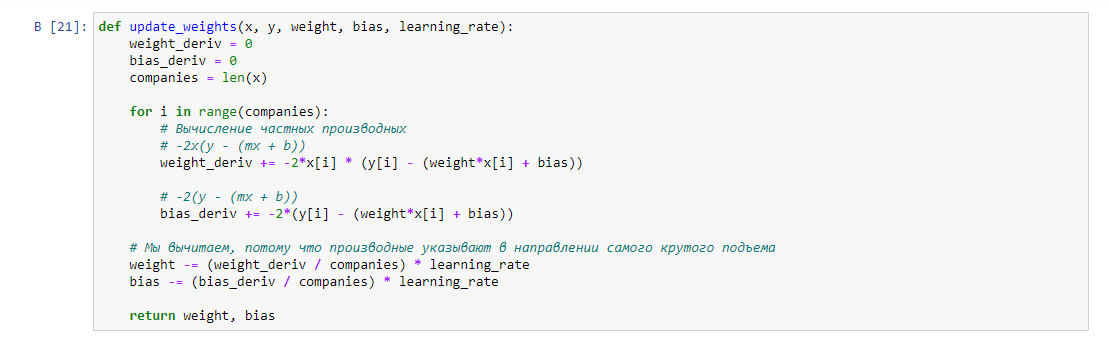
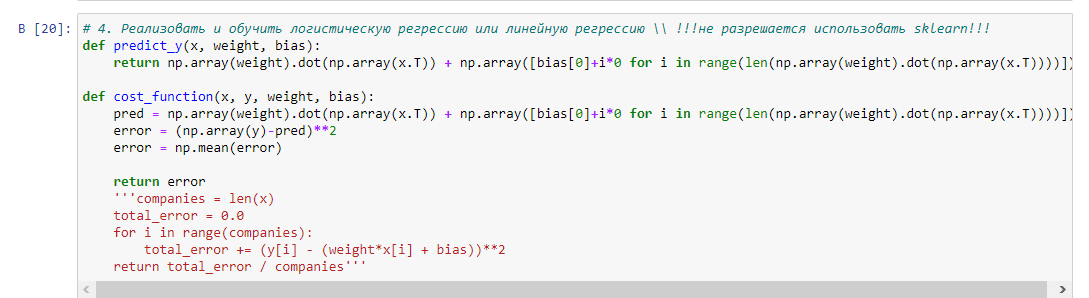


Далее мы разделяем данные на обучающую и тестовую выборку, а также нормализуем X-данные с помощью StandardScaler.

Реализуем метрику качества:

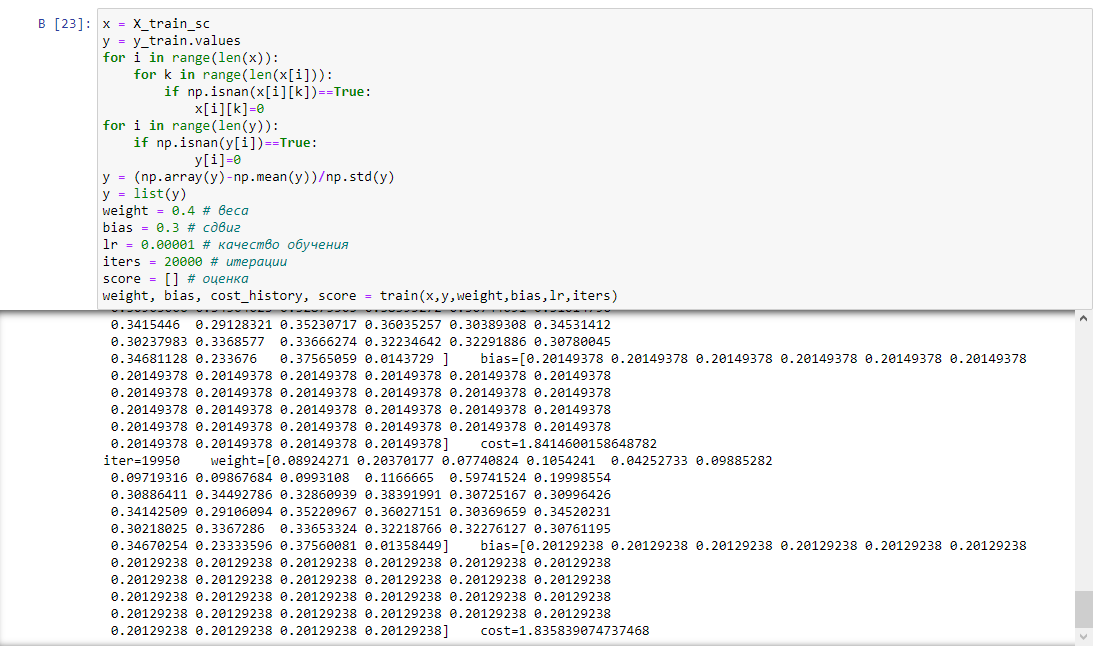


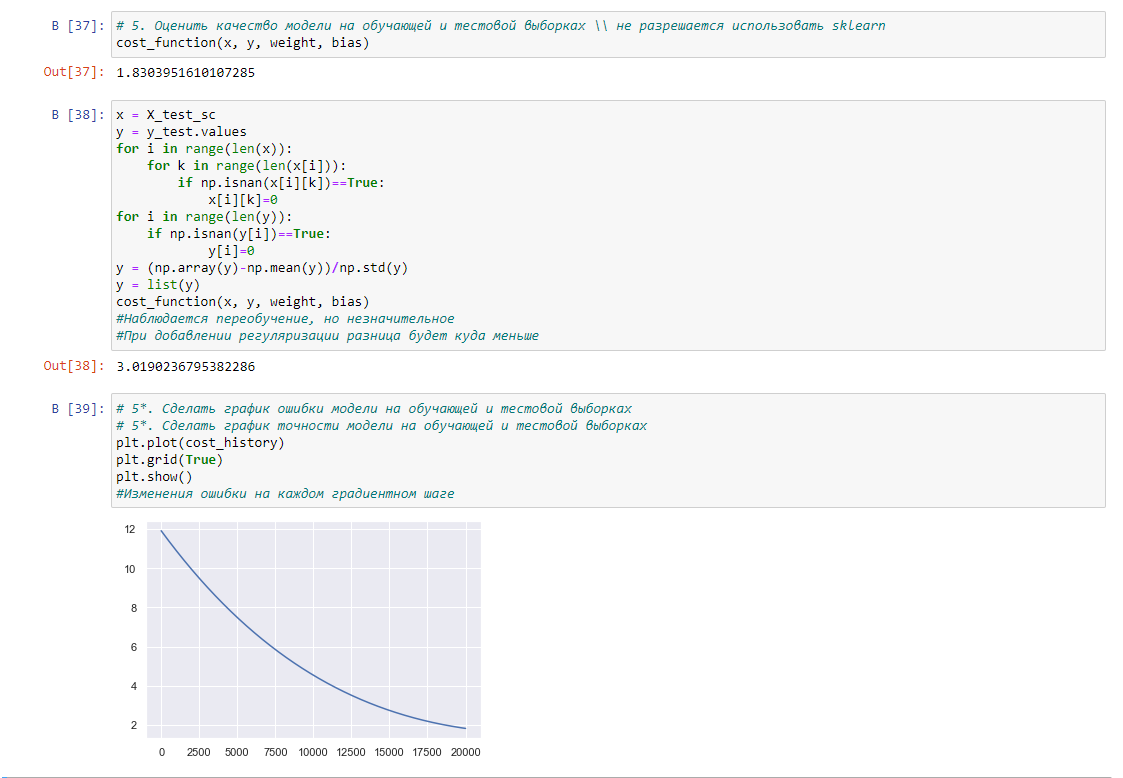
Реализуем линейную регрессию:

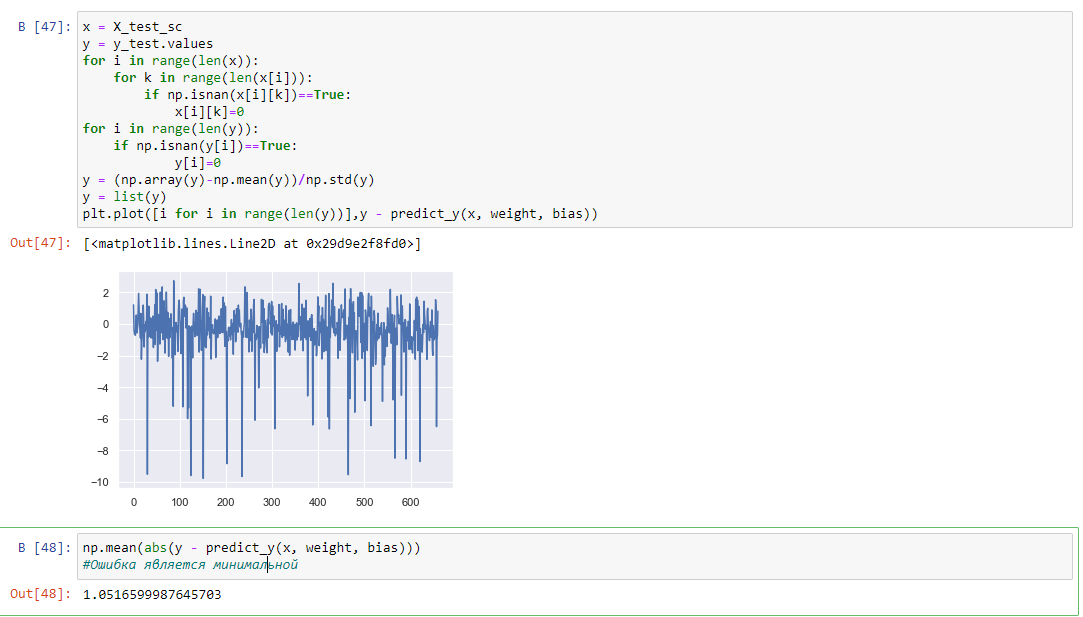
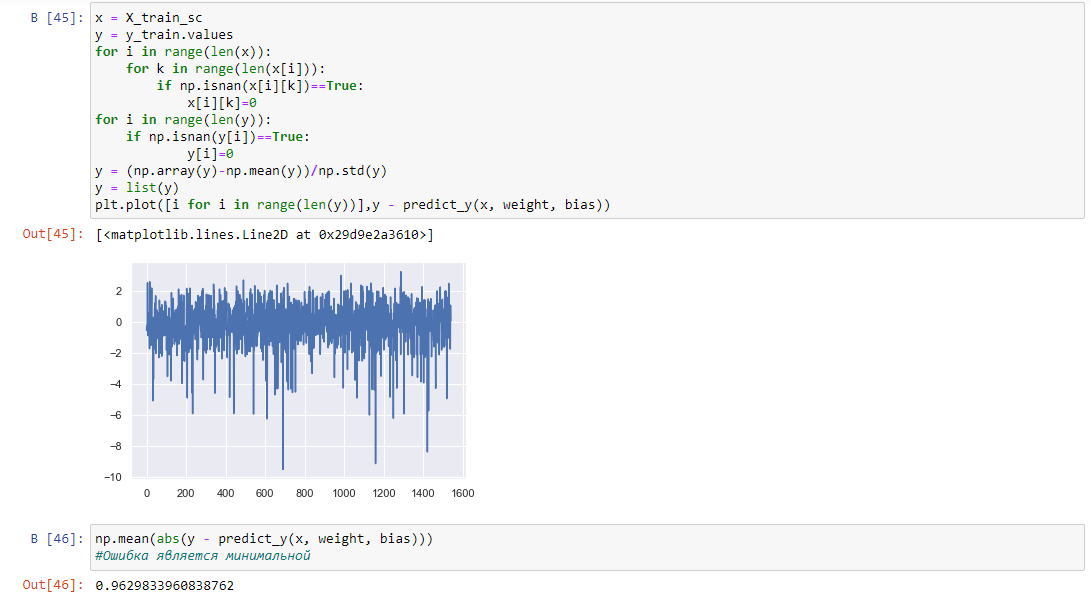




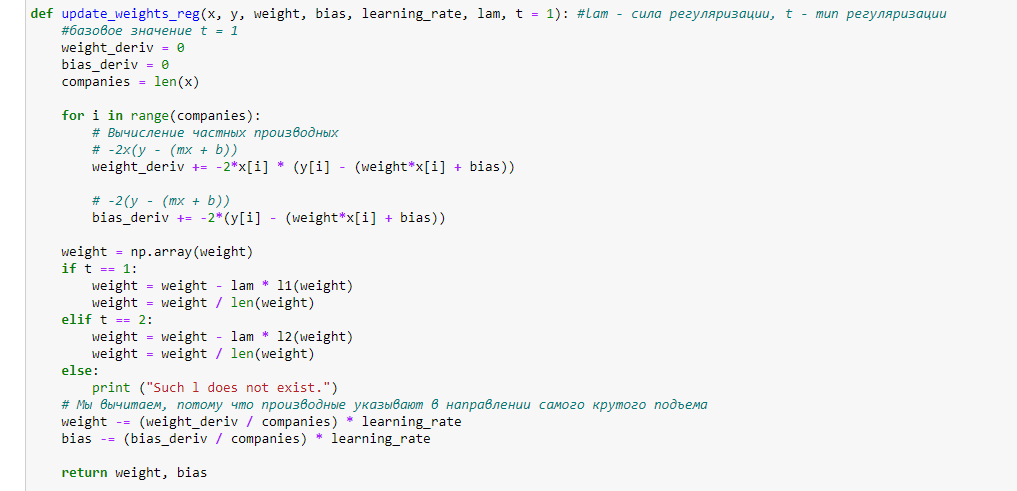
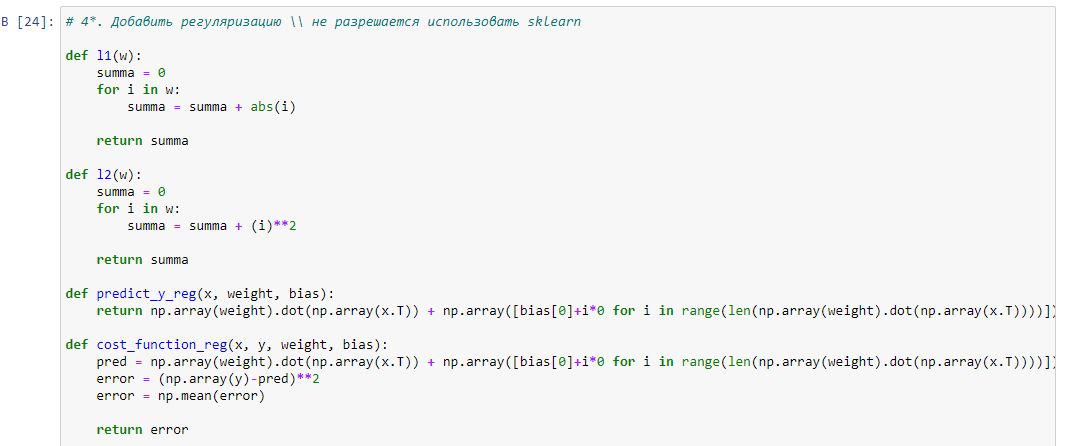
Обучим линейную регрессию:

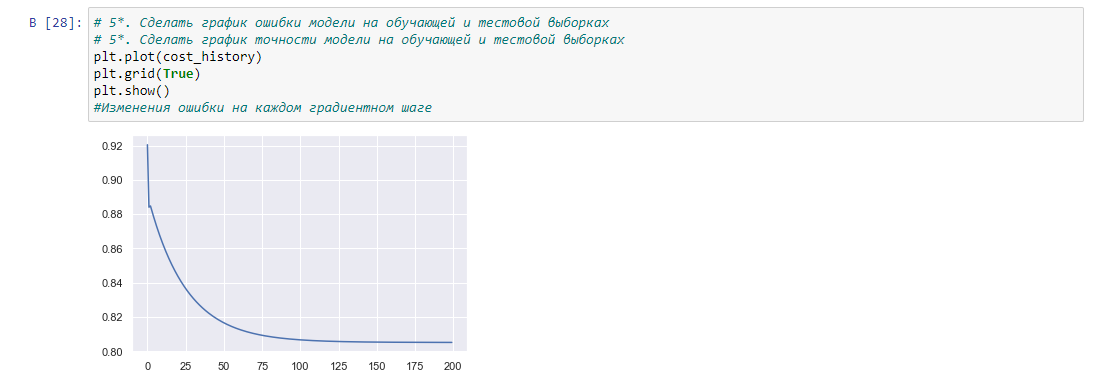
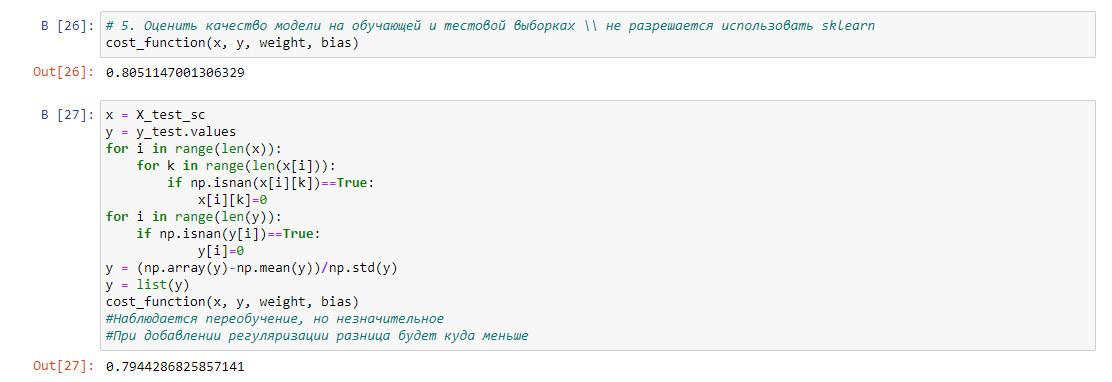
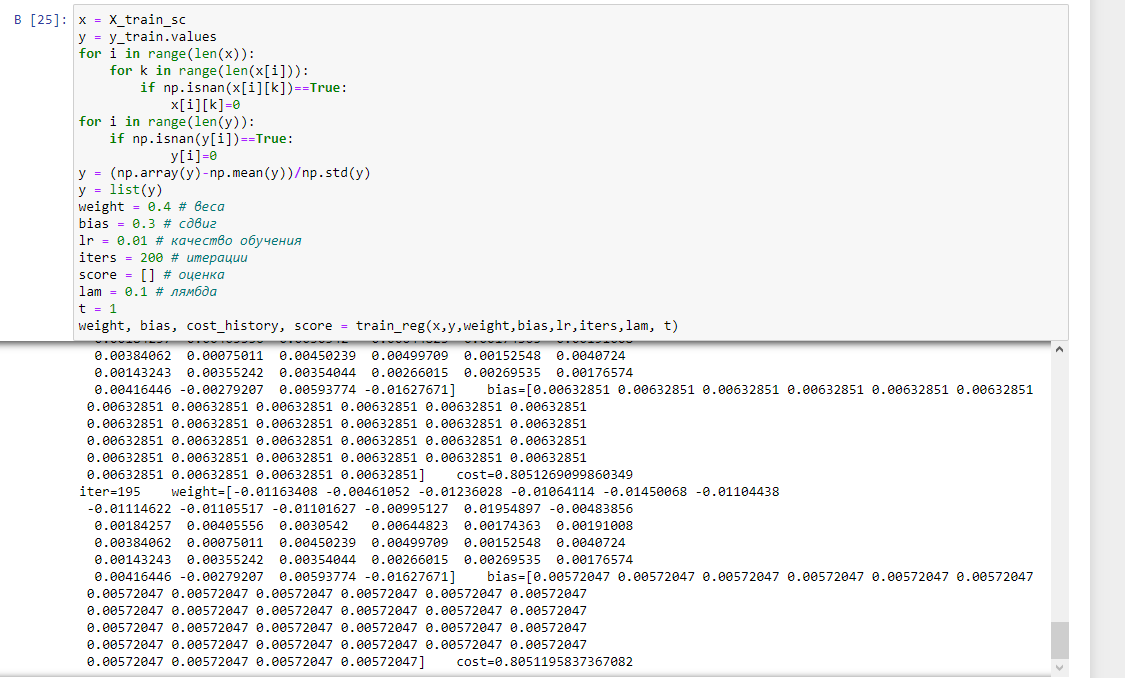


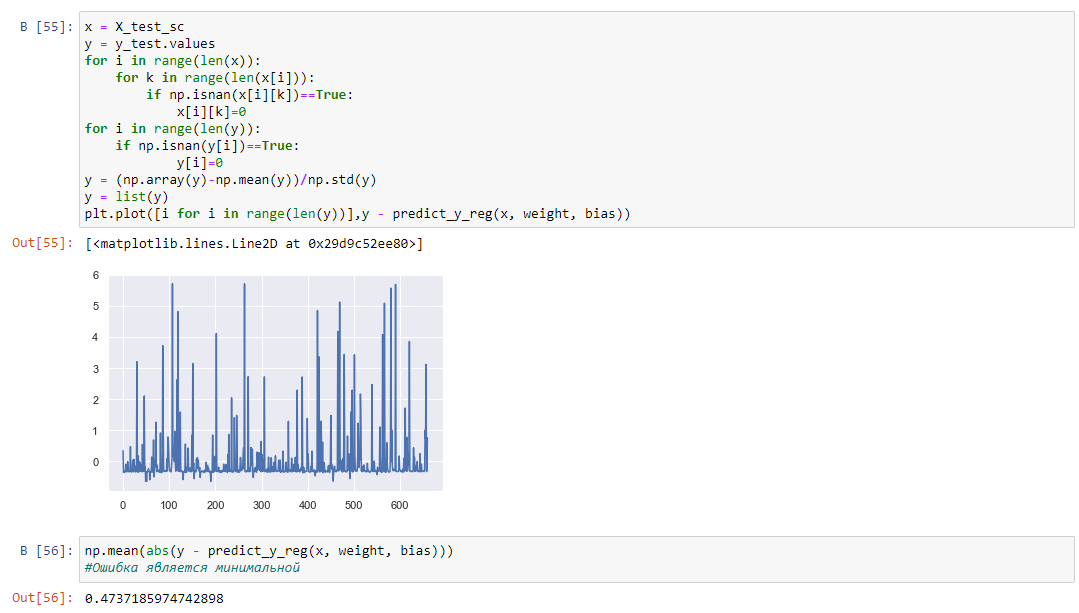
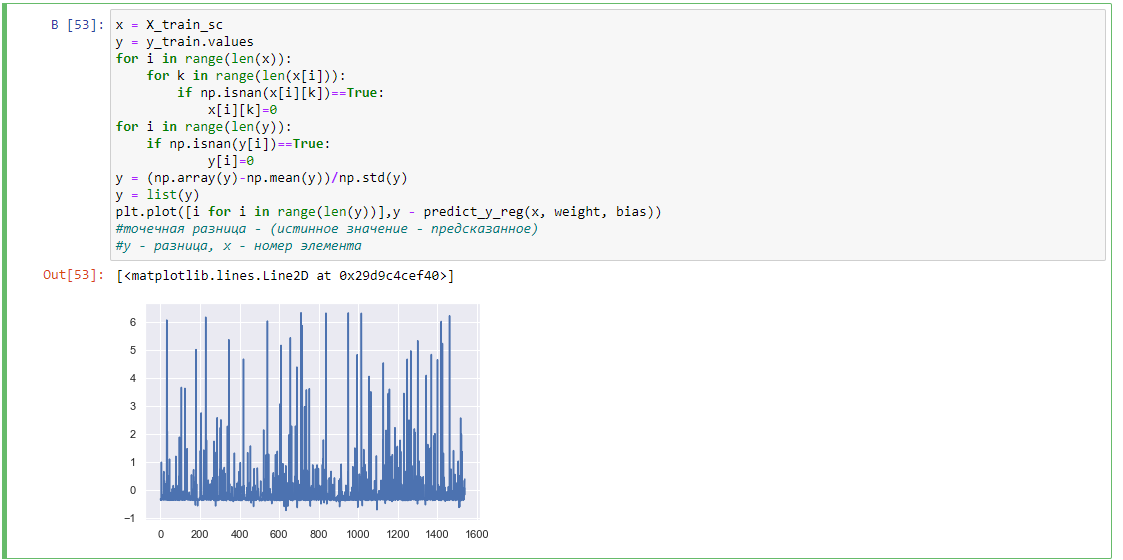




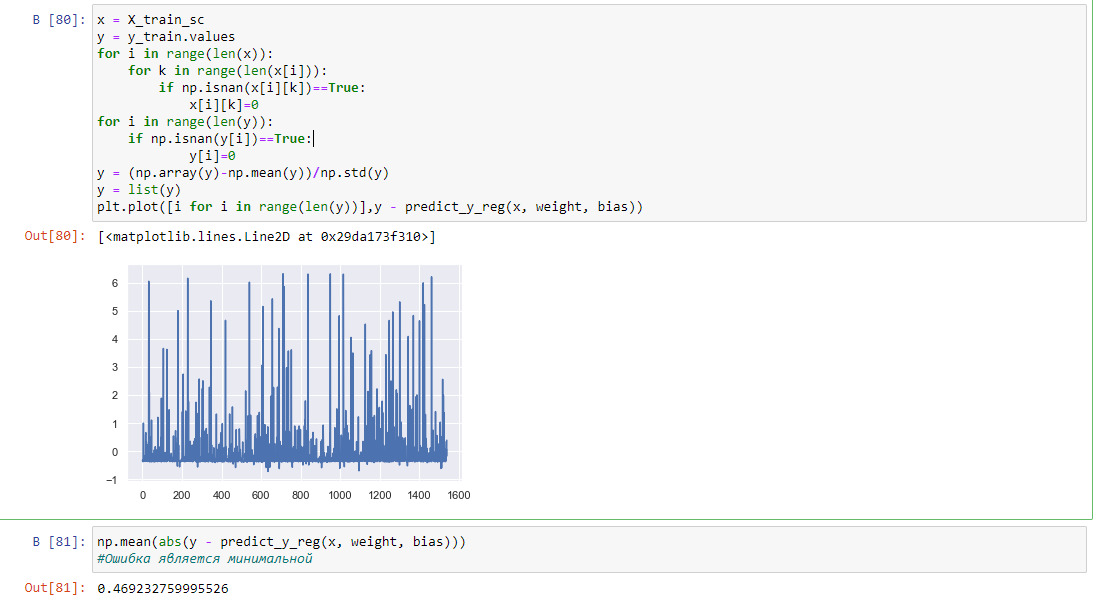
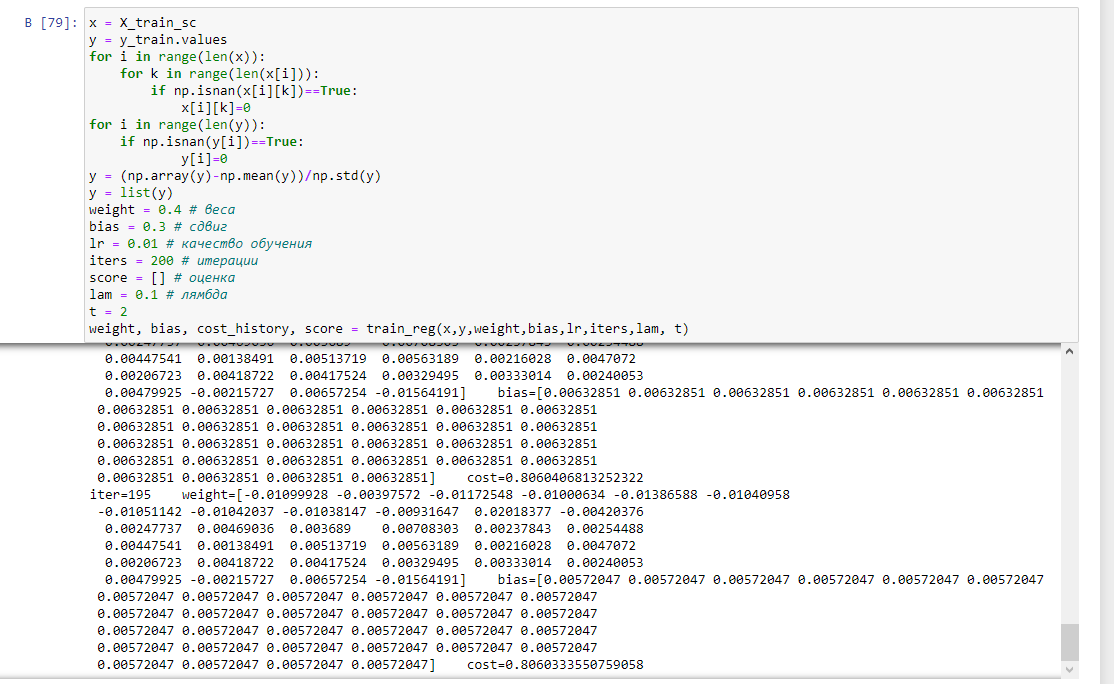
Из результатов наблюдаем, что у нас есть небольшое переобучение, поэтому мы производим регуляризацию нашего градиента

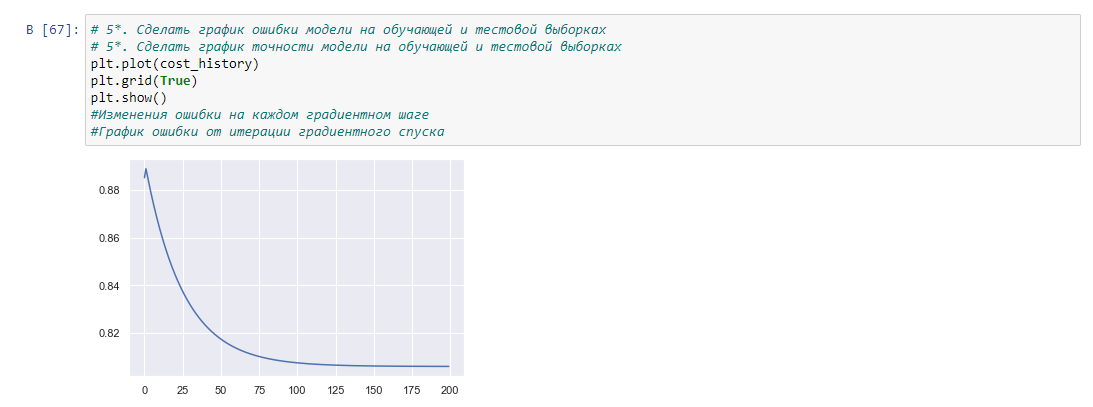
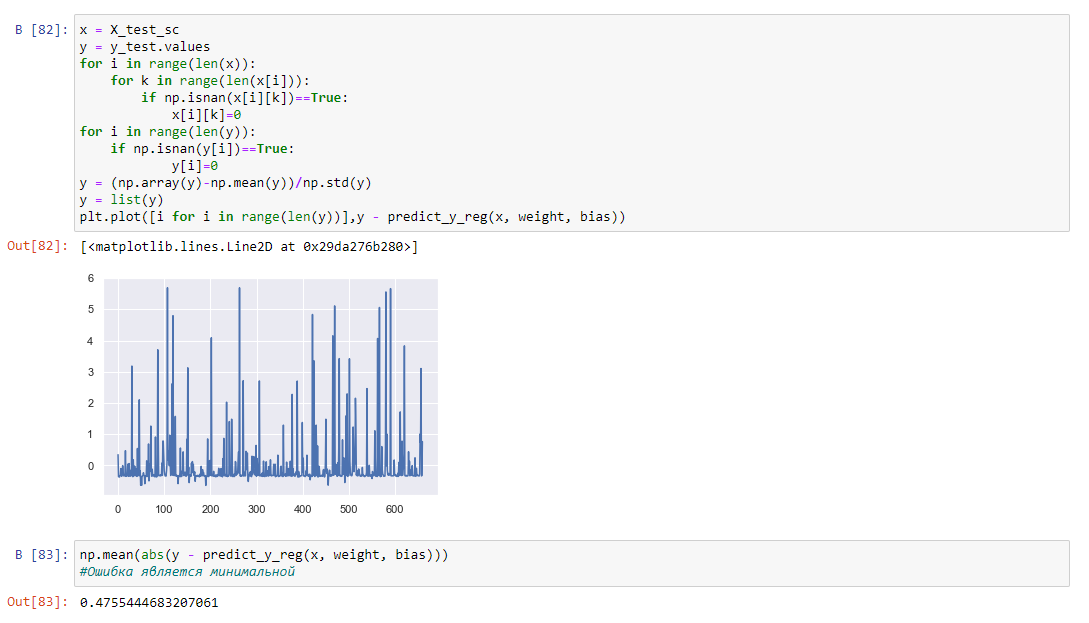


Обучаем регулизацию L1:



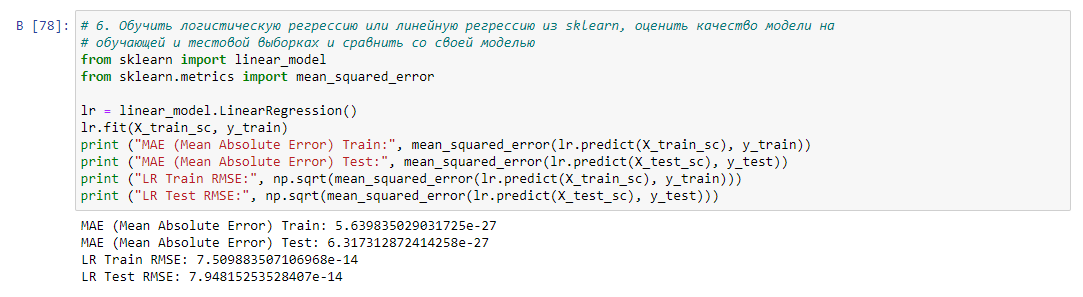
Обучаем регулизацию L2:

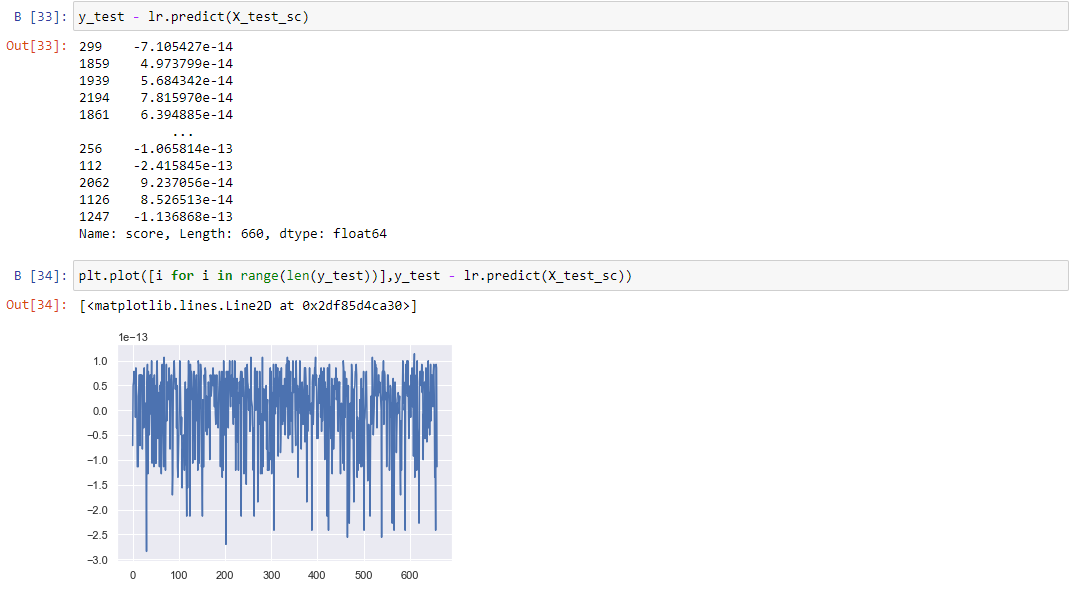
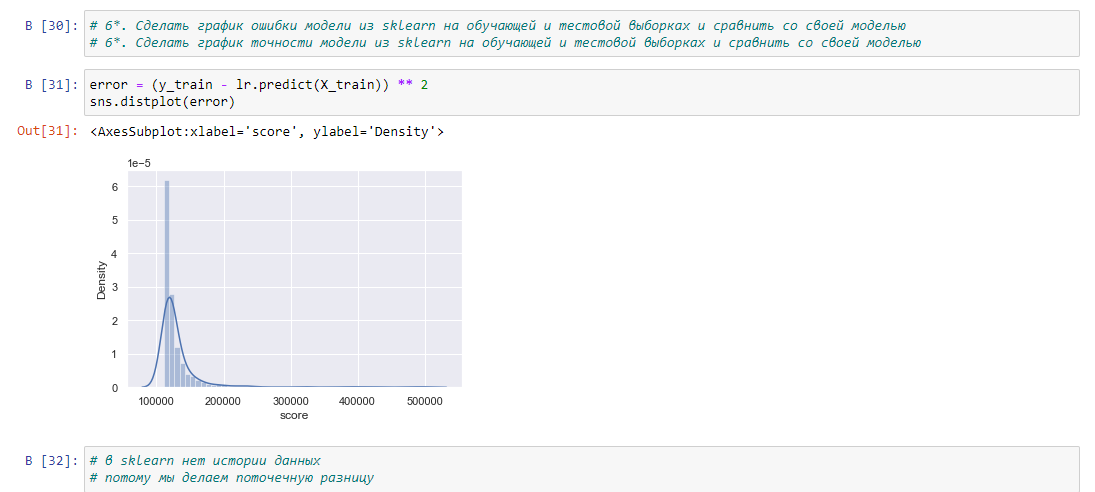


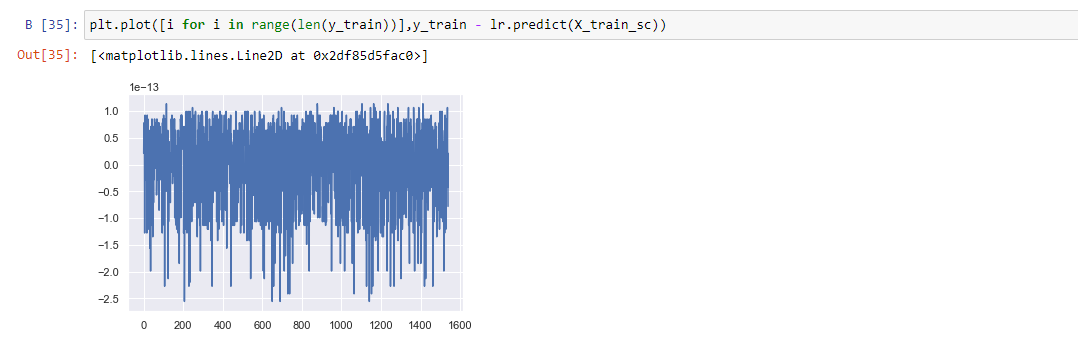


Нам удалось устранить переобучение.

Теперь сделаем то же самое, но с помощью sklearn:





Посмотрим качество моделей по ошибкам

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ЛР№4 (без регуляризации) | КП (с регуляризацией L1) | КП (с регуляризацией L2) | SKLearn |
| обучающая | 1.8303951610107285 | 0.8051147001306329 | 0.8060284714698305 | 7.509883507106968e-14 |
| тестовая | 3.0190236795382286 | 0.7944286825857141 | 0.7937310307734364 | 7.94815253528407e-14 |

**Вывод**

Познакомился с Линейной регрессией, которая является простым, но мощным механизмом для аппроксимации линейных зависимостей.

Модель является предшественником нелинейных методов, которые используют для обучения нейронных сетей. В ходе курсового проекта удалось улучшить реализацию модели, добавить регуляризацию, и добиться лучших оценок.