МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Московский государственный технический университет имени Н.Э. Баумана

(национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science»

Слушатель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (Полунина М.М.)

Москва, 2023

**Содержание**

|  |  |  |
| --- | --- | --- |
|  | Введение …………………………………………..………………………... | 3 |
| 1. | Аналитическая часть ……………………...…………………..…………… | 4 |
| 1.1 | Постановка задачи ………………………………………………..………... | 5 |
| 1.2 | Описание используемых методов………………………………..………... | 7 |
| 1.3 | Разведочный анализ данных, визуализация основных показателей……. | 13 |
| 2. | Практическая часть……………………………………………………..….. | 20 |
| 2.1 | Предобработка данных, подготовка к созданию модели……………....... | 21 |
| 2.2 | Разработка, обучение и тестирование моделей…………………………… | 27 |
| 2.3 | Создание рекомендательной системы…………………………………..… | 32 |
|  | Заключение ………………………………………………………………..... | 35 |
|  | Приложение ……………………………………………………………….... | 36 |
|  | Источники ………………………………………………………………….. | 37 |

**Введение**

Каждый день человечество генерирует примерно 2,5 квинтиллиона байт различных данных. Они создаются буквально при каждом клике и пролистывании страницы, не говоря уже о просмотре видео и фотографий в онлайн-сервисах и соцсетях.

Наука о данных появилась задолго до того, как их объемы превысили все мыслимые прогнозы. Отсчет принято вести с 1966 года, когда в мире появился Комитет по данным для науки и техники — CODATA. Его создали в рамках Международного совета по науке, который ставил своей целью сбор, оценку, хранение и поиск важнейших данных для решения научных и технических задач. В составе комитета работают ученые, профессора крупных университетов и представители академий наук из нескольких стран, включая Россию.

Развитие Data Science шло вместе с внедрением технологий Big Data и анализа данных. И хотя эти области часто пересекаются, их не следует путать между собой. Все они предполагают понимание больших массивов информации. Но если аналитика данных отвечает на вопросы о прошлом (например, об изменениях в поведениях клиентов какого-либо интернет-сервиса за последние несколько лет), то Data Science в буквальном смысле смотрит в будущее.

Область применения анализа данных достаточно широка. Реализация технологий Data Science не обошла стороной и медицинскую сферу. Ярким примером является использование университетом Флориды свободных данных из больниц общественного здравоохранения и карты Google для создания визуальных данных, которые позволяли бы быстрее идентифицировать и эффективнее анализировать медицинскую информацию, для отслеживания распространения хронических заболеваний.

Анализ данных также применяется для улучшения ряда аспектов жизнедеятельности городов и стран. Например, зная точно, как и когда использовать технологии Data Science, можно оптимизировать потоки транспорта. Для этого берется в расчет передвижение автомобилей в режиме онлайн, анализируются социальные медиа и метеорологические данные. Сегодня ряд городов взял курс на использование анализа данных с целью объединения транспортной инфраструктуры с другими видами коммунальных услуг в единое целое. Это концепция «умного» города, в котором автобусы ждут опаздывающий поезд, а светофоры способны прогнозировать загруженность на дорогах, чтобы минимизировать пробки.

Фондовый рынок не стал исключением для применения технологий Data Science. Комиссия по ценным бумагам (SEC) использует большие данные для мониторинга деятельности фондового рынка. В настоящее время они используют сетевую аналитику и обработку естественного языка, чтобы поймать незаконную торговую деятельность на фондовых рынках.

Розничные трейдеры, крупные банки и другие на фондовых рынках используют большие данные для анализа торговли, используемые в высокочастотной торговле, аналитике поддержки принятия решений до торговли, измерении настроений, прогнозной аналитике и т. д.

Технологии Data Science позволили разработать более сложные алгоритмы прогнозирования и контролировать их эффективность в реальном времени. Сбор, фильтрация, углубленный анализ с использованием сложных статистических моделей и удобная визуализация огромного объема данных, генерируемых фондовым рынком, создают конкурентное преимущество для инвесторов, использующих современные технологии в формировании успешных торговых стратегий.

В книжном издательском бизнесе, как и в любой другой предпринимательской сфере, также накапливаются данные. Это не только рейтинги книг и количество продаж, но и данные по тиражности изданий, популярности определенных жанров, количестве языков, статистика реализованных экземпляров классическим способом – через книжные магазины, а также в онлайн формате (интернет-магазины). И всё это в рамках нашей большой страны – тоже Big Data.

Российская Книжная палата предоставляет немало интересных статистических отчетов касаемо книжной отрасли, но данных которые можно было бы проанализировать самостоятельно в открытом доступе, к сожалению, нет. Поэтому было принято решение провести исследование открытого набора данных с ресурса kaggle.com под названием Goodreads-books, который представляет собой список книг с определенными рейтинговыми оценками, годом выпуска книги, издательством и другими показателями.

На основе представленного набора была написана данная выпускная квалификационная работа, состоящая из введения, двух глав и заключения.

В первой главе осуществлена постановка задачи данной работы, проблематика отрасли в целом, описаны основные используемые методы для решения задачи и проведен разведочный анализ.

Вторая глава состоит из практической части исследования: загрузки и предобработки данных, создания моделей машинного обучение, их тестирование и сравнение. Во вторую главу также входят данные по созданию рекомендательной системы.

В заключении представлены выводы по результатам исследования предложенного датасета.

**1.1 Постановка задачи**

Выпуск и реализация книжной продукции точно также поддается анализу, как и любые другие товары, выпускаемые производителями на рынок. Анализ современной издательской системы позволяет детализировать основные показатели развития книгоиздания, оценить динамику выпуска книг и брошюр по отдельным сегментам и тиражным группам, выпуск переводной литературы. Детально изучения деятельность крупнейших игроков издательского рынка и рейтинг ведущих издательских домов мира позволяют количественные показатели отрасли в целом, инфраструктура книгораспространения, данные, характеризующие экспорт и импорт книжной продукции и др.

К сожалению, подробно оценить вышеперечисленные параметры отечественного книжного рынка не является возможным, т.к. в открытом доступе представлены уже готовые статистические и аналитические отчеты по данному вопросу и найти именно исходные данные для исследования не удалось.

Поэтому было принято решение провести исследование открытого набора данных из источника www.kaggle.com под названием Goodreads-books, который представляет собой список книг с сайта книжного сообщества www.goodreads.com c определенными рейтинговыми оценками, годом выпуска книги, издательством и другими показателями.

В данной работе была поставлена задача регрессионного анализа: исследование зависимости среднего рейтинга от всех остальных категориальных и не категориальных признаков.

Одновременно с этим необходимо осуществить создание рекомендательной системы, основанной на классификации рейтингов книжных изданий и показать возможности использования этой системы.

**1.2 Описание используемых методов**

Если под Big Data мы будем понимать не сами данные, а методы их обработки, которые позволяют эффективно работать с информацией, то суть этих методов следует рассмотреть подробнее.

В данной выпускной квалификационной работе широко использованы методы EDA – Exploratory Data Analysis, разведочный анализ данных, благодаря которому удобно оценить полноту предоставленных данных, проанализировать основные показатели и визуально представить их при помощи графиков. Разведочный анализ данных— один из первых и определяющих шагов проекта науки о данных, который приводит в движение весь проект. Он придает проекту конкретное направление и формирует план его реализации.

Разведочный анализ данных означает изучение данных до самых глубин для получения из них практической информации. Он включает в себя анализ и обобщение массивных наборов данных, часто в форме диаграмм и графиков. Основная цель разведочного анализа – получить полное понимание о данных и привлечь внимание к их наиболее важным особенностям. Такой анализ позволяет подготовить данные к применению в алгоритмах машинного обучения. Кроме того, он помогает генерировать гипотезы о данных, выявлять аномалии и структуру.

Корреляционный анализ в свою очередь также играет немаловажную роль в исследованиях. Благодаря ему становится возможным оценить, есть ли зависимость между показателями и какую переменную необходимо объявить целевой для построения регрессионной модели машинного обучения. Корреляция – это статистическая зависимость между случайными величинами, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой. Корреляционный анализ – это раздел математической статистики, посвященный изучению взаимосвязей между случайными величинами. Корреляционный анализ заключается в количественном определении тесноты связи между двумя признаками (при парной связи) и между результативным и множеством факторных признаков (при многофакторной связи). Коэффициент корреляции принимает значения от -1 до +1. Положительное значение коэффициента корреляции свидетельствует о наличии прямой связи, отрицательное – обратной.

Данный метод обработки статистических данных весьма популярен в экономике, астрофизике и социальных науках (в частности в психологии и социологии), хотя сфера применения коэффициентов корреляции обширна: контроль качества промышленной продукции, металловедение, агрохимия, гидробиология, биометрия и прочие. В различных прикладных отраслях приняты разные границы интервалов для оценки тесноты и значимости связи.

Популярность метода обусловлена двумя моментами: коэффициенты корреляции относительно просты в подсчете, их применение не требует специальной математической подготовки. В сочетании с простотой интерпретации, простота применения коэффициента привела к его широкому распространению в сфере анализа статистических данных.

Существует три распространенных способа измерения корреляции:

* Корреляция Пирсона: используется для измерения корреляции между двумя непрерывными переменными. (например, рост и вес). Коэффициент корреляции Пирсона оценивает только линейную связь переменных. Нелинейную связь данный коэффициент выявить не может.
* Корреляция Спирмена: используется для измерения корреляции между двумя ранжированными переменными (например, оценка балла учащегося на экзамене по математике и оценка на экзамене по естественным наукам в классе). Метод ранговой корреляции Спирмена позволяет определить тесноту (силу) и направление корреляционной связи между двумя признаками или двумя профилями (иерархиями) признаков
* Корреляция Кендалла: используется в том случае, если нет возможности использовать корреляцию Спирмена – размер выборки слишком мал. Как и коэффициент ранговой корреляции Спирмена, коэффициент Кендалла используется для измерения взаимосвязи между качественными признаками, характеризующими объекты одной и той же природы, ранжированные по одному и тему же критерию. Изменяется от -1 до +1. Коэффициенты корреляции Спирмена и Кендалла используются как меры взаимозависимости между рядами рангов, а не как меры связи между самими переменными.

Опираясь на разведочный анализ данных необходимо всегда перед построением регрессионных моделей решать вопрос об очистке и возможном синтезировании данных. Поскольку в приведенном наборе Goodrears-books почти полностью отсутствовали пропущенные значения и проблема выбросов была не критичной, работы по предварительной подготовке данных для создания модели была минимальной.

В данной работе была поставлена задача регрессионного анализа – выявление того, насколько изменение одной или нескольких переменных в среднем влияет на изменение другой переменной (результативного признака).

Для проведения регрессионного анализа были использованы такие методы, как:

* LinearRegression
* AdaBoostRegressor (DecisionTreeRegressor)
* RandomForest Regressor

Линейная регрессия (Linear regression) — используемая в статистике регрессионная модель зависимости одной (объясняемой, зависимой) переменной **y** от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) **x** с линейной функцией зависимости.

Модель линейной регрессии является часто используемой и наиболее изученной в эконометрике. А именно изучены свойства оценок параметров, получаемых различными методами при предположениях о вероятностных характеристиках факторов, и случайных ошибок модели. Предельные (асимптотические) свойства оценок нелинейных моделей также выводятся исходя из аппроксимации последних линейными моделями. С эконометрической точки зрения более важное значение имеет линейность по параметрам, чем линейность по факторам модели.

Ансамблевая модель AdaBoost (Adaptive Boosting) реализует принципы так называемого «бустинга», используя неглубокие случайные деревья DecisionTree класса Regressor для последовательного обучения, а затем уменьшения ошибок в выборке, учитывая предыдущий «опыт». Бустинг приобрел свою популярность благодаря возможности увеличения алгоритмов, входящих в модель. Алгоритм AdaBoostRegressor устойчив к выбросам.

Алгоритм случайного леса Random Forest является ярким представителем другого ансамблевого метода машинного обучения - «бутстрэпа». Суть Бутстрэпа заключается в формировании определенного количества новых обучающих выборок на основе одной исходной. Для этого случайным образом выбирается k-й элемент выборки и копируется в новую (в прежней он остается, не удаляется). Затем, эта операция повторяется m раз – по заданному размеру новой выборки. Таким образом формируется новая обучающая выборка, состоящая из элементов исходной с некоторыми повторениями, т.к. вполне возможно несколько раз случайно отобрать один и тот же элемент. После формирования одной выборки, алгоритм переходит к формированию следующей и так T раз для T выборок, которые, очевидно, будут несколько отличаться друг от друга.

Метод случайных подпространств, используемый в «бутстрэпе», позволяет снизить коррелированность между деревьями и избежать переобучения. Базовые алгоритмы обучаются на различных подмножествах признакового описания, которые также выделяются случайным образом. Класс алгоритма RandomForest Regressor используется в машинном обучении для решения задач регрессии.

Модели регрессии используются для количественной оценки взаимосвязи между одной или несколькими переменными-предикторами и переменной отклика. При подборе регрессионной модели важно понять, насколько хорошо модель может использовать значения переменных-предикторов для прогнозирования значения переменной отклика.

Все регрессионные алгоритмы, используемые в данной работе, оцениваются по двум параметрам: MAE (средняя абсолютная ошибка) и R2 (показатель детерминации).

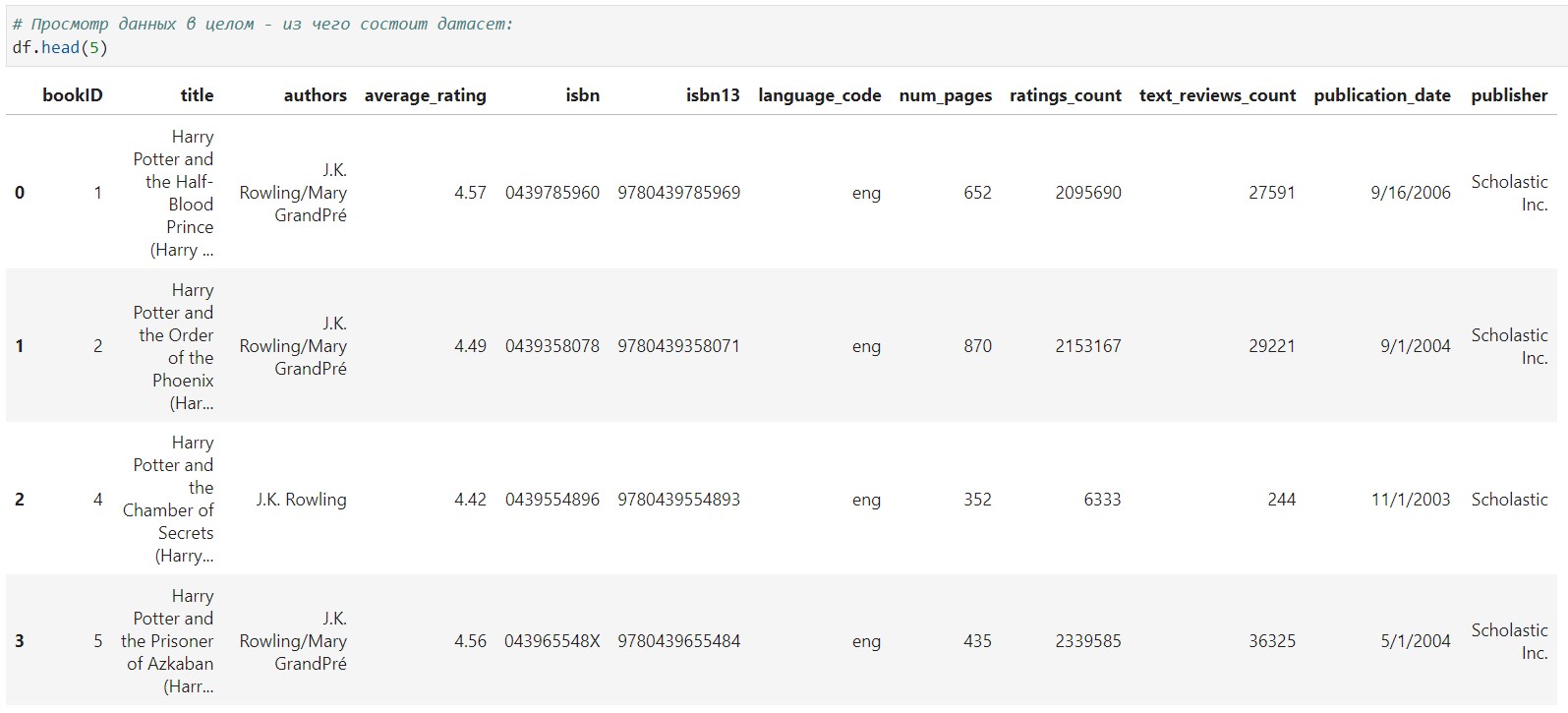
* MAE : самая распространенная метрика функции потерь, которая сообщает среднюю абсолютную разницу между прогнозируемыми значениями и фактическими значениями в наборе данных. MAE не предает большое значение выбросам и является линейной оценкой, а это значит, что все ошибки в среднем взвешены одинаково. Чем ниже MAE, тем лучше модель соответствует набору данных.
* R2 : это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными. Его рассматривают как универсальную меру зависимости одной случайной величины от множества других. В частном случае линейной зависимости R2 является квадратом так называемого множественного коэффициента корреляции между зависимой переменной и объясняющими переменными. В частности, для модели парной линейной регрессии коэффициент детерминации равен квадрату обычного коэффициента корреляции между **y** и **x**.

За выбор лучшего сценария предсказания моделей отвечает функция кросс-валидации GridSearchCV. Кросс-валидация – очень важная техника в машинном обучении (применяемая также в статистике и эконометрике), с ее помощью выбираются гиперпараметры моделей, сравниваются модели между собой, оценивается полезность новых признаков в задаче и т.д. GridSearch осуществляет поиск лучших параметров в фиксированной сетке возможных значений.

Для создания базовой рекомендательной системы в данной работе используется KNN (k-nearest neighbors algorithm) – метод ближайших соседей. KNN это метрический алгоритм для автоматической классификации объектов или регрессии. В случае использования метода для классификации объект присваивается тому классу, который является наиболее распространённым среди k соседей данного элемента, классы которых уже известны. В случае использования метода для регрессии, объекту присваивается среднее значение по k ближайшим к нему объектам, значения которых уже известны.

Алгоритм может быть применим к выборкам с большим количеством атрибутов (многомерным). Для этого перед применением нужно определить функцию расстояния; классический вариант такой функции — Евклидова метрика.

**1.3 Разведочный анализ данных, визуализация основных показателей**

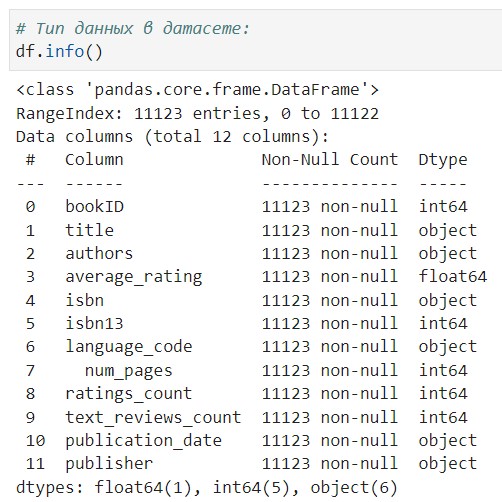
Набор данных из источника [www.kaggle.com](http://www.kaggle.com) Goodreads-books состоит из 11123 строк и 12 столбцов (рисунок 1).

*Рисунок 1 – Первые строки набора данных Goodreads-books*

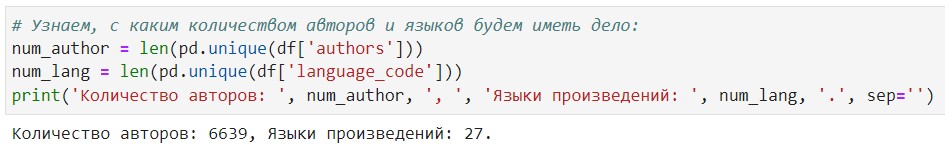
Из набора данных можно увидеть, что здесь осуществлен подсчет рейтинга книг интернет ресурса [www.goodreads.com](http://www.goodreads.com), где пользователи самостоятельно составляют рейтинг выпущенных книг публикуя текстовые отзывы или просто проставляя рейтинг произведениям.

Подробнее о наименовании столбцов:

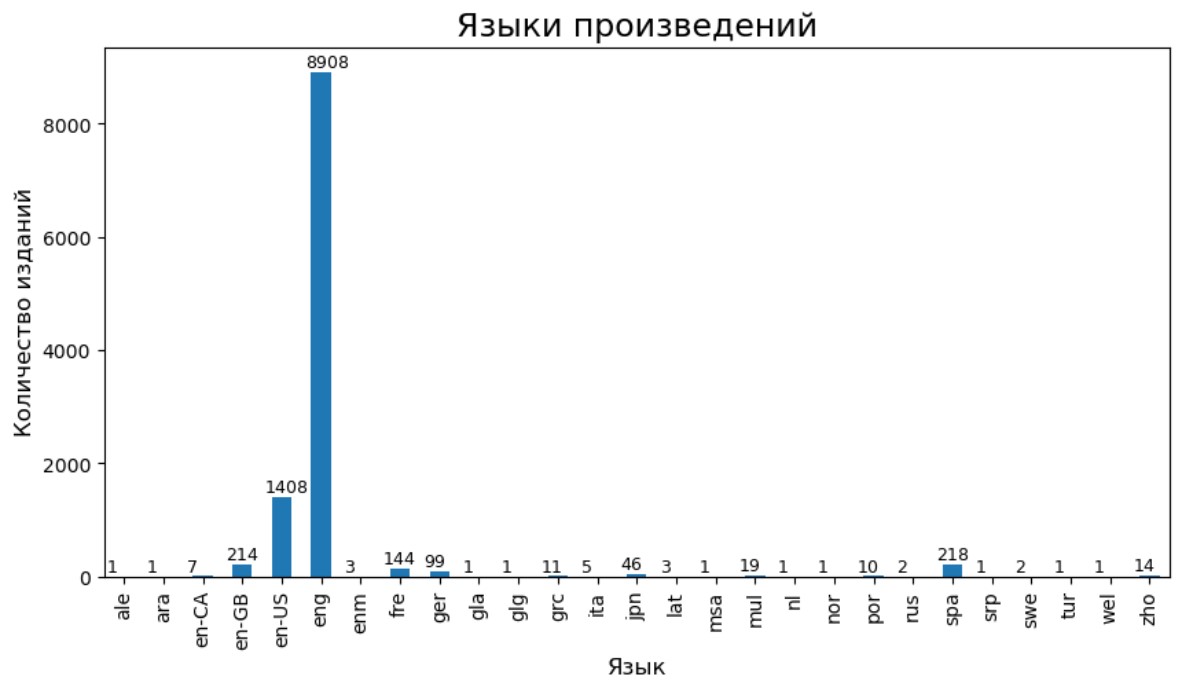
* 'bookID' - Номер книги по порядку в данном датасете;
* 'title' - Название книги;
* 'authors' - Имя/Фамилия автора книги;
* 'average\_rating' - Средний рейтинг;
* 'isbn' - Международный стандартный книжный номер (10 цифр);
* 'isbn13' - Международный стандартный книжный номер (13 цифр);
* 'language\_code' - Язык издания;
* 'num\_pages' - Количество страниц;
* 'ratings\_count' - Количество учтенных рейтингов;
* 'text\_reviews\_count' - Количество текстовых отзывов;
* 'publication\_date' - Дата выхода книги в печать;
* 'publisher' - Издательство, выпустившее книгу.

Разведочный анализ данных позволяет проанализировать структуру данных, предоставленных в наборе.

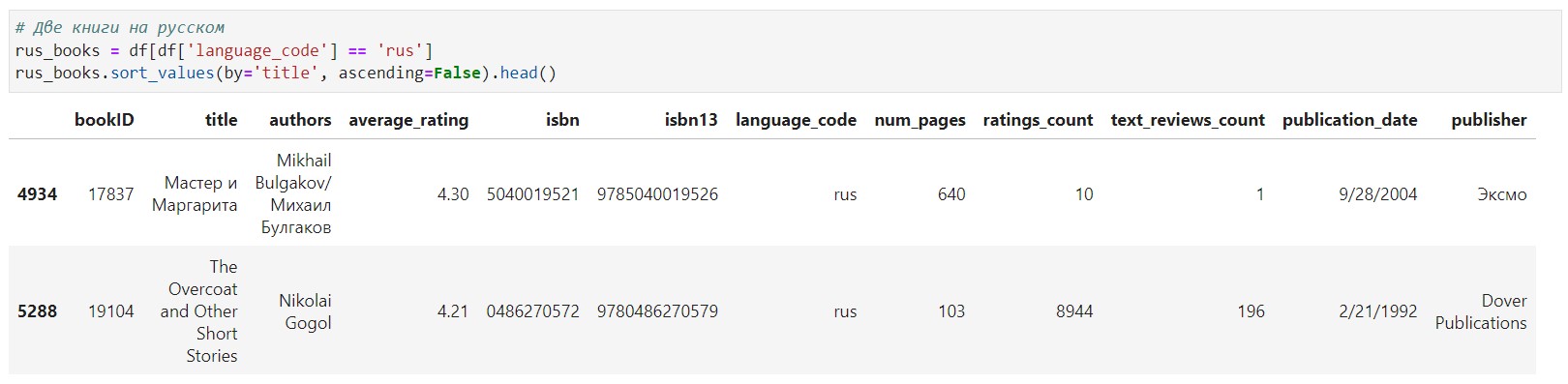
*Рисунок 2 – Типы данных в датасете*

Посмотреть сколько всего языков и авторов произведений можно при помощи небольшой команды.

*Рисунок 3 – Общее количество авторов и языков*

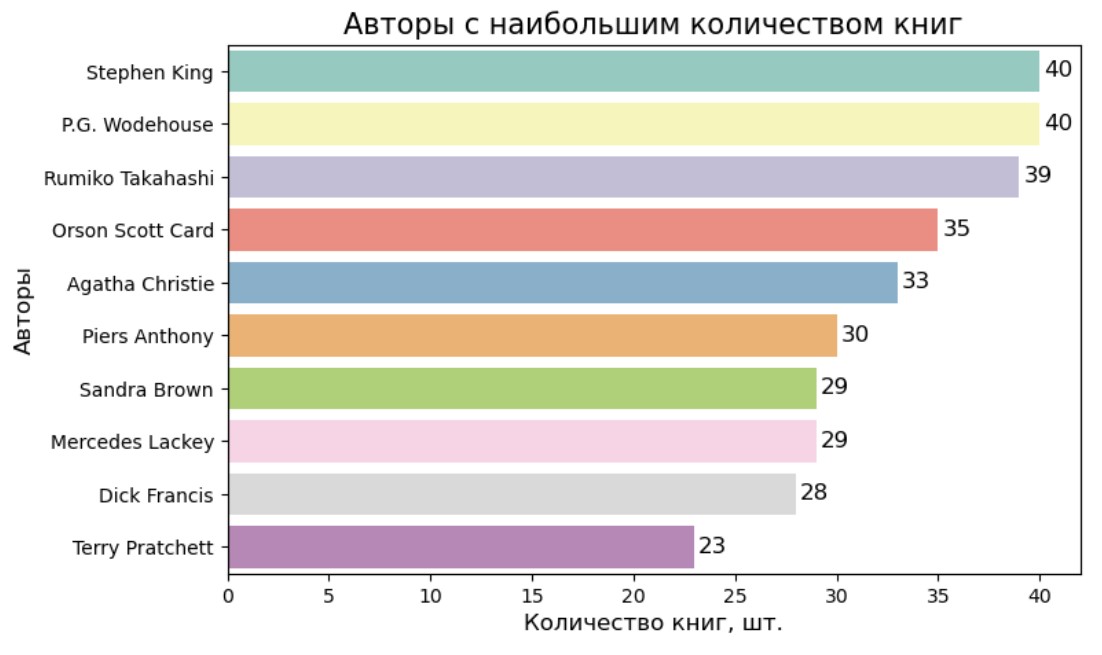
Наглядно количество языков представлено в виде столбчатого графика на Рисунке 4.

*Рисунок 4 – Языки произведений*

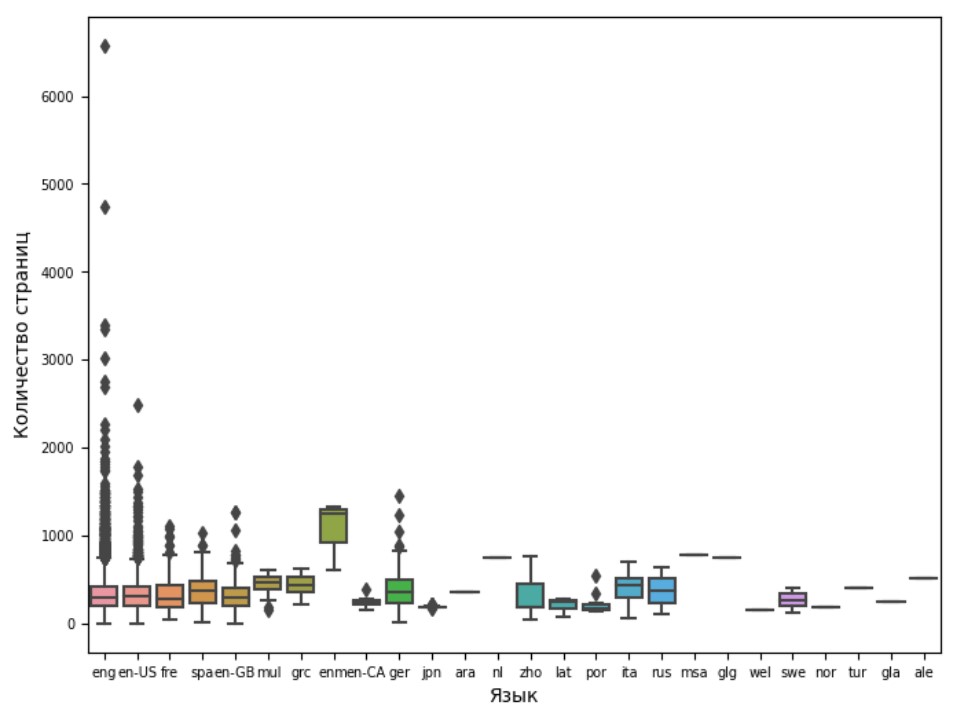
Среди большого количества книг нашлись две на русском языке: «Шинель» Н.В. Гоголя и «Мастер и Маргарита» М.Ю. Булгакова (рисунок 5).

*Рисунок 5 – Книги на русском языке*

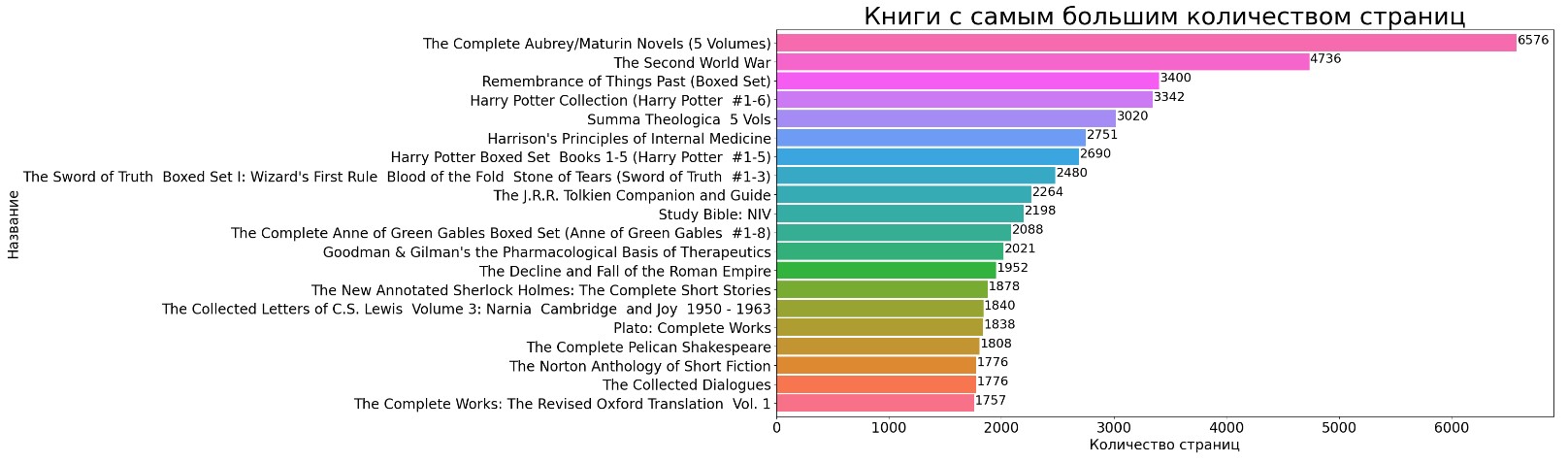
*Рисунок 6 – Топ 10 языков по количеству книг*

На рисунке 6 можно обратить внимание, что в основной массе преобладают книги на английском языке, причем категориальный признак явно один, но имеет несколько похожих значений.

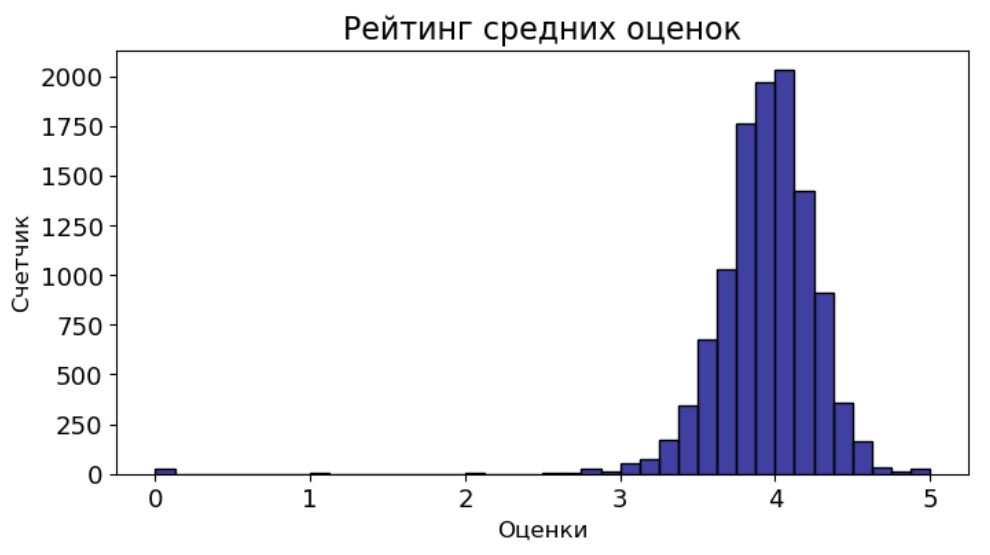
*Рисунок 7 – Авторы с наибольшим количеством книг*

Рисунок 7 носит ознакомительный характер, позволяя понять какие авторы имеют самое большое количество книг в этом наборе данных.

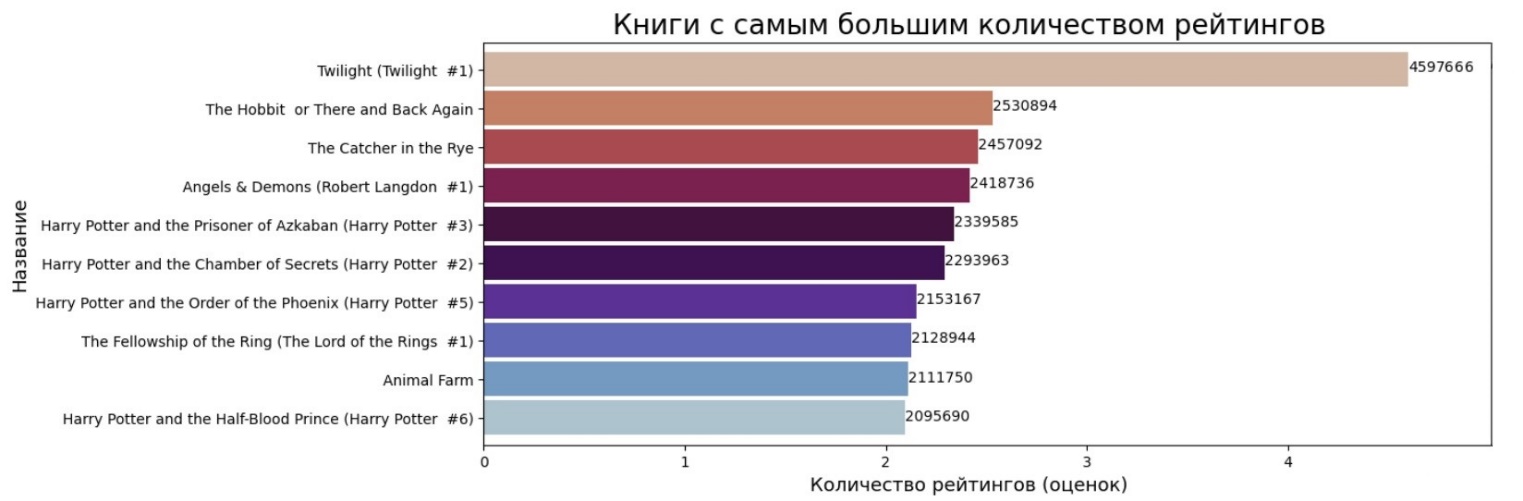
*Рисунок 8 –График выбросов языкового признака относительно количества страниц*

Возвращаясь к теме языков на рисунке 8 представлен график выбросов, так называемый «Ящик с усами», для признака «Язык» относительно признака «Количество страниц» из которого следует, что количество явных выбросов в этом наборе данных минимально, но в дальнейшем небольшая чистка всё же потребуется.

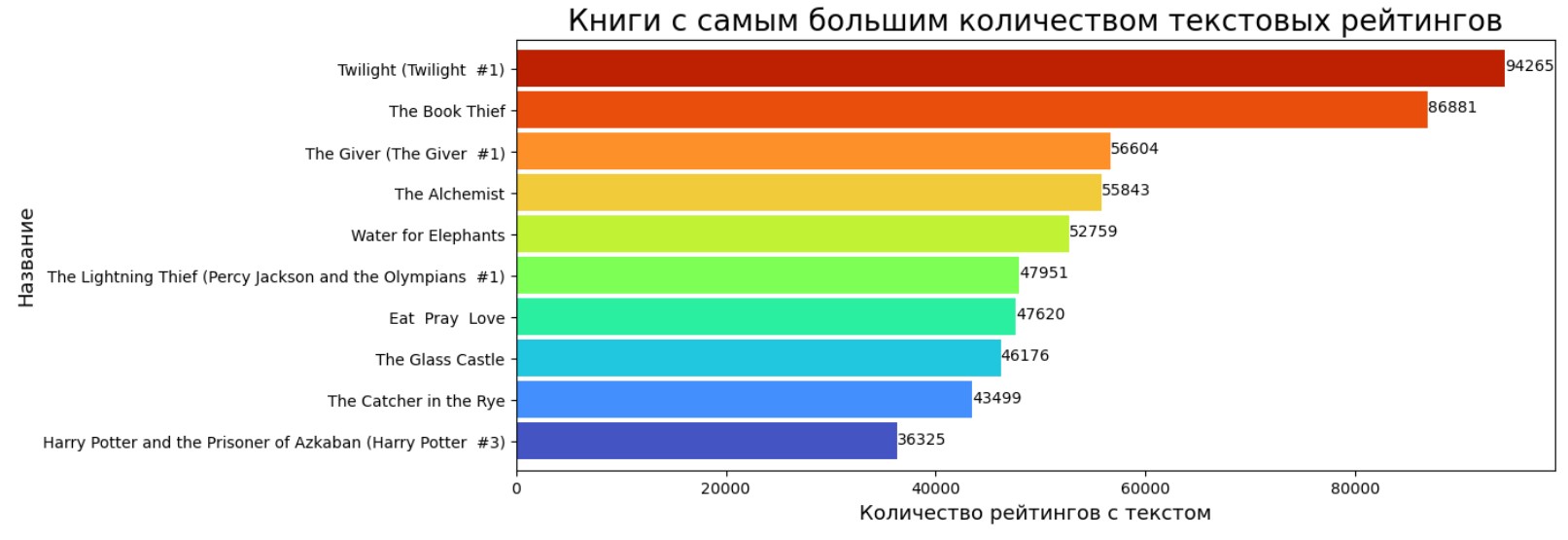
*Рисунок 9 – Книги с самым большим количеством страниц*

Рисунок 9 наглядно иллюстрирует, что именно из себя представляют выбросы: книги с наибольшим количеством страниц это многотомники, собрания сочинений и тп.

*Рисунок 10 – Рейтинг средних оценок*

В наборе данных Goodreads-books основная масса оценок находится в диапазоне приблизительно от 3.8 до 4.3 баллов в среднем. Количество оценок ниже 2.0 минимально, но какие-то данные всё же присутствуют.

*Рисунок 11 – Книги с самым большим количеством рейтингов (оценок)*

На рисунках 11 и 12 детально показано, на какие произведения чаще всего обращали внимания пользователи сервиса, какие книги чаще всего оценивали, а каким – аще писали отзывы с текстом.

*Рисунок 12 – Книги с самым большим количеством текстовых рейтингов*

Посмотрев таким образом на данные, имеющиеся в этом наборе, можно увидеть, что в целом данные очищены, полны и особых преобразований не требуют. Но всё же потребуется немного предварительной работы по удалению выбросов и небольшие преобразования языковых признаков.

**2.Практическая часть**

В практической части данной работы предстоит провести немало манипуляций: подготовка данных планируется минимальная, но всё же её необходимо сделать.

После этапа подготовки данных следует разработка регрессионных моделей машинного обучения, которые последовательно будут обучаться на тестовой части данных, а затем проходить тестирование на валидационной части.

Планируется привести сравнительные данные по каждому из примененных методов. Необходимо будет проанализировать, как отработала каждая модель и почему дала определенные результаты.

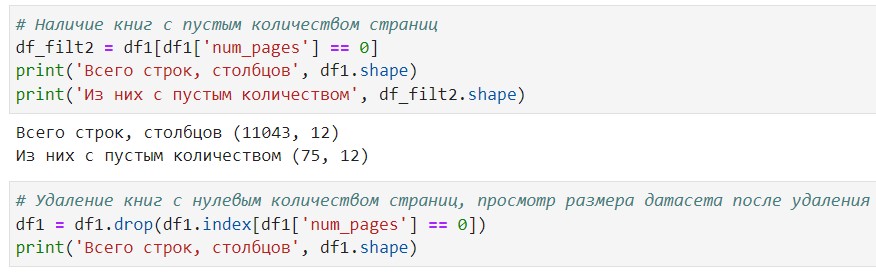
Отдельным пунктом будет разработана рекомендательная система, подбор книг в которой должен осуществляться исходя из рейтинга определенного произведения.

**2.1 Предобработка данных, подготовка к созданию модели**

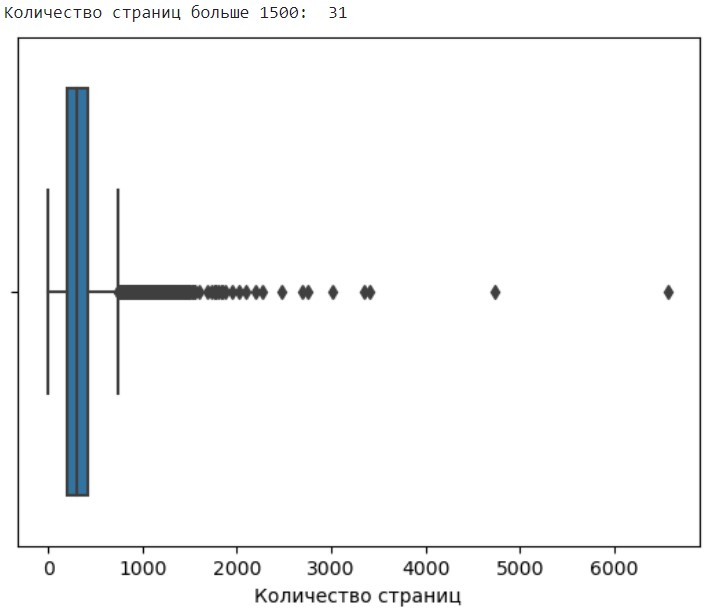
Чтобы иметь возможность независимо поработать с данными в первую очередь необходимо сделать копию. Именно на скопированном наборе данных и будет происходить предобработка.

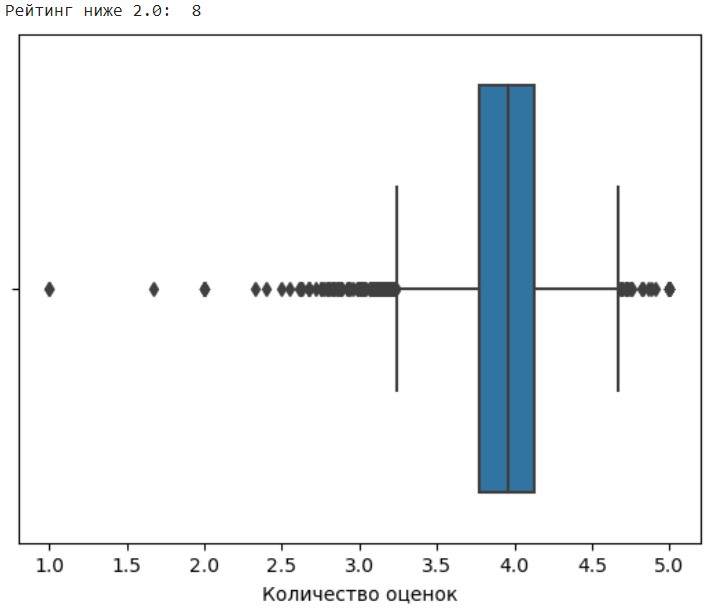
****Для начала необходимо удалить книги, у которых рейтинг не проставлен (равен 0):

*Рисунок 13 – Обнаружение и удаление книг без рейтинга*

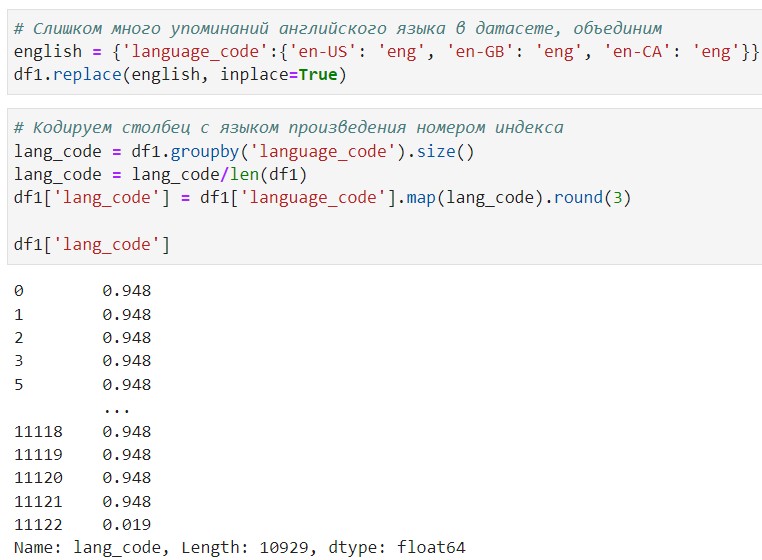
Следующим этапом необходимо избавиться от значений книг, где есть рейтинги и отзывы, но количество страниц указано как 0. Эти данные не пригодятся для построения моделей.

*Рисунок 14 – Обнаружение и удаление книг без страниц*

**Для повышения точности работы моделей также необходимо уделить внимание выбросам, а именно удалить из набора данных книги с очень большим количеством страниц и с рейтингом ниже 2.0.

*Рисунок 15 – «Ящик с усами» для книг, с количеством страниц больше 1500*

*Рисунок 16 – «Ящик с усами» для книг, с рейтингом меньше 2.0*

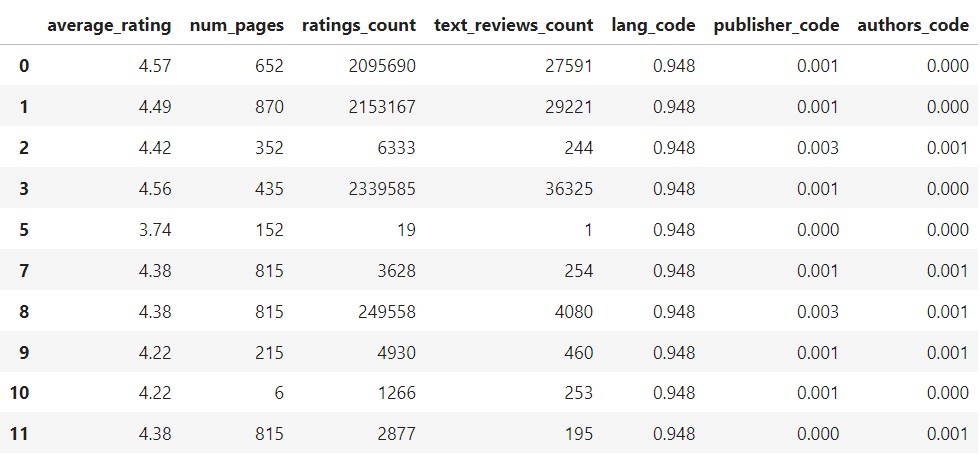
Предстоит немного поработать с категориальным признаком «Язык произведения», в котором присутствуют некоторые повторы в наименовании языка. В рамках работы с этим столбцом можно сразу закодировать значения числом, которое отражает количество индексов данного значения в столбце.

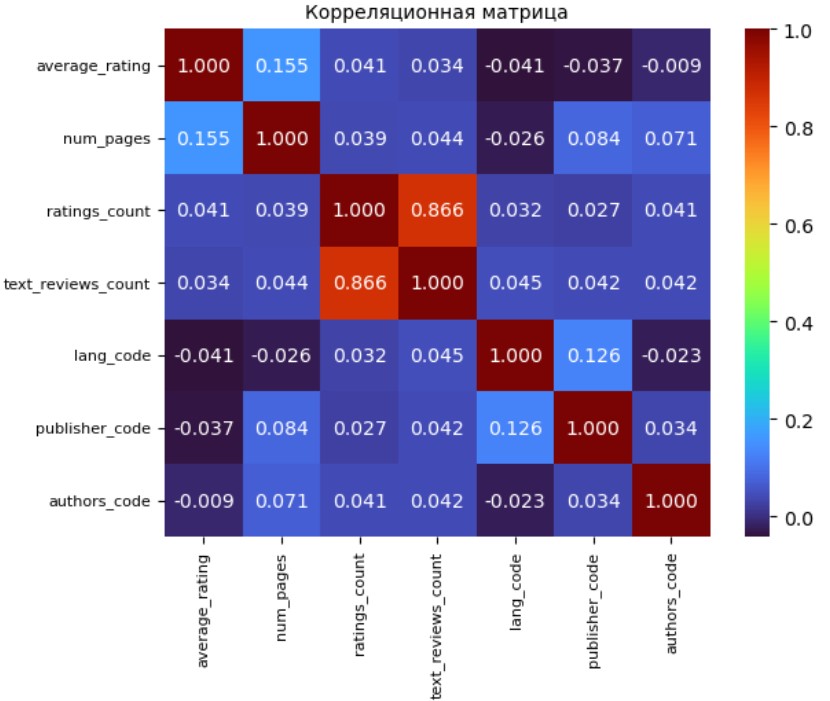
*Рисунок 17 – Манипуляции со столбцом «Язык произведения»*

Аналогичным образом кодируются и другие столбцы с категориальными признаками: наименование издательства, выпустившего книгу и имя автора произведения.

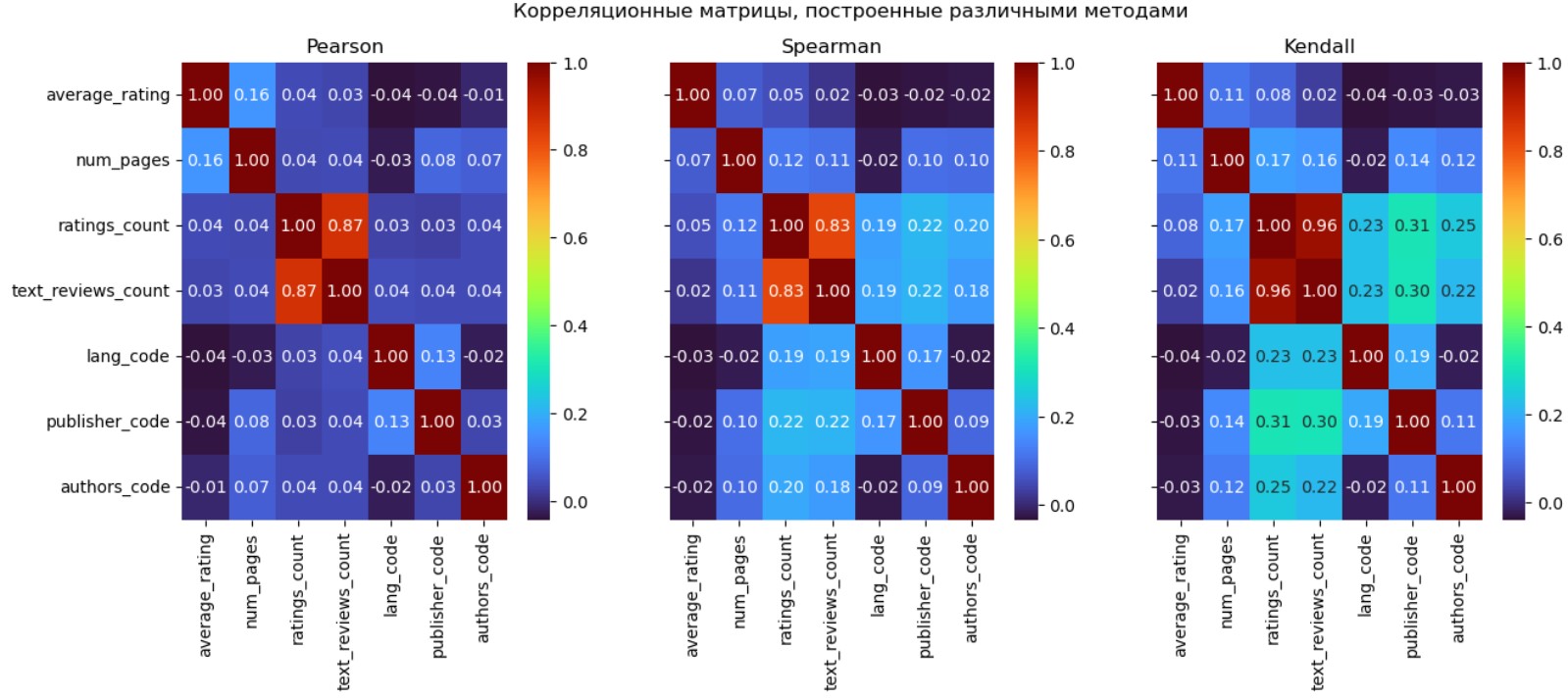
После проверки результатов кодирования и удаления столбцов, значения которых носят уникальный характер (таких как «BookID», «IBSN», «IBSN13»), а также столбцов-дубликатов набор данных приобретает вид, с которым уже можно работать, строя регрессионные модели.

Вид подготовленных и очищенных данных представлен на рисунке 18.

*Рисунок 18 – Очищенные данные*

Очистка данных и подготовка их к работе завершена, но перед тем как перейти к построению регрессионных моделей необходимо понять, что будет целевым признаком и насколько данные коррелируют друг с другом в этом наборе.

*Рисунок 19 – Корреляционная матрица*

На рисунках 19 и 20 представлены корреляционные матрицы, которые показывают, насколько данные в этом наборе зависят или не зависят друг от друга.

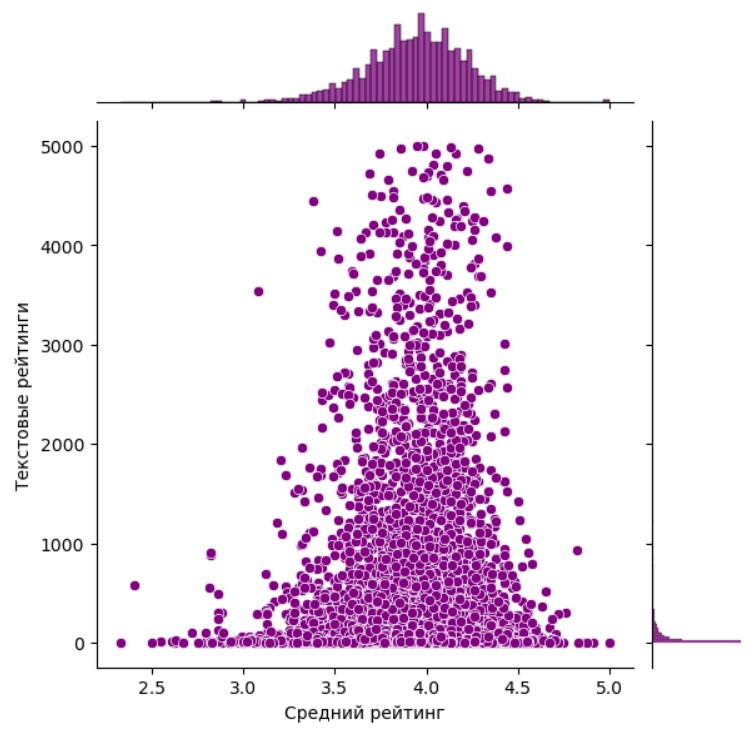
*Рисунок 20 – Корреляционные матрицы, построенные различными методами*

На основе корреляционных матриц можно сделать следующие выводы:

Прослеживается взаимосвязь между рейтингами, содержащими текст и общим количеством учтенных рейтингов. Можно сделать предположение, что при наличии текста в оценке книги её средний рейтинг становится выше. Есть смысл поискать, какие рейтинги с текстом наиболее высокие и у каких авторов.

****Соответственно, если у одного автора много высокорейтинговых книг, можно предположить, что в будущем продажи его книг будут на очень высоком уровне.

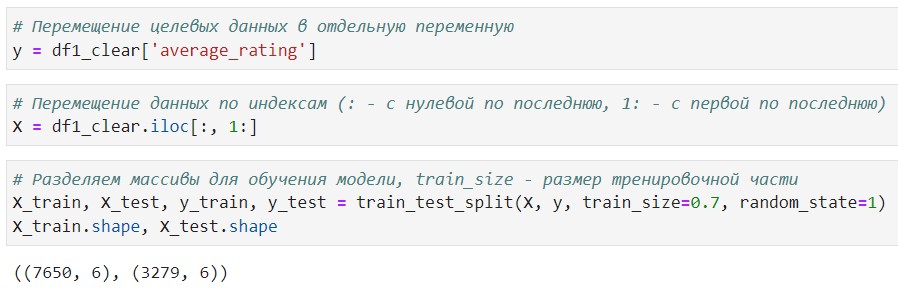
*Рисунок 21 – Корреляция среднего рейтинга к остальным показателям*

Совершенно очевидно, что на этом наборе данных целевой переменной станет «Средний рейтинг», предсказание которого сможет в целом показывать насколько будущие книги определенных авторов будут популярны у аудитории.

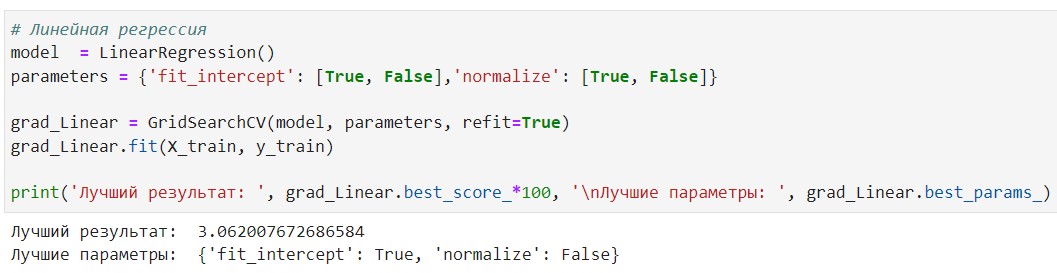
*Рисунок 22 – Комбинированный график среднего рейтинга к текстовым рейтингам*

Подготовка и анализ данных на этом этапе закончены, самое время переходить к построению моделей.

**2.2 Разработка, обучение и тестирование моделей**

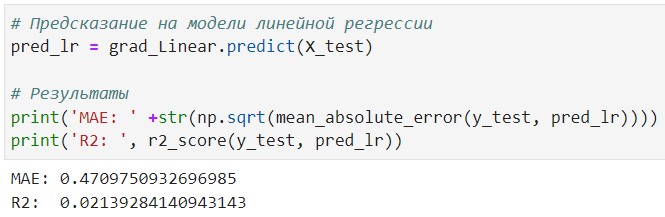
В первую очередь необходимо отделить от общего набора данных целевую переменную **y**, во вторую – поместить всё остальное. Затем подготовленные данные необходимо разделить на две части: тренировочную часть, на котороый модель будет обучаться и тестовую, где будет возможность проверить качество обучения модели.

*Рисунок 23 – Подготовка данных для подачи в модель*

Первая модель, которой будут «скормлены» данные – LinearRegression, линейная регрессия.

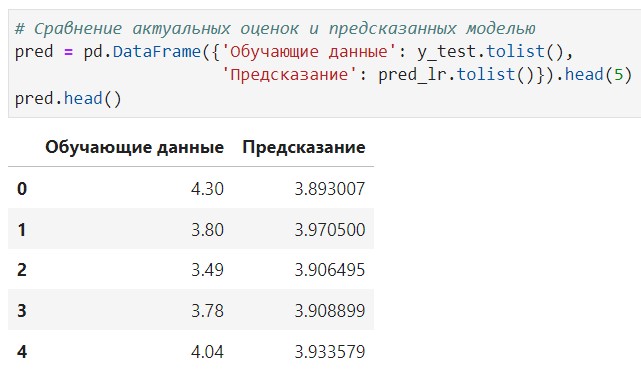
*Рисунок 24 – Модель линейной регрессии*

Поскольку в работу модели добавлен инструмент кросс-валидации GridSearchCV после отработки тестовых данных модель показывает, каким был лучший результат и на каких лучших параметрах.

Сразу после отработки модели на тренировочных данных, можно применить обученный алгоритм к тестовым данным.

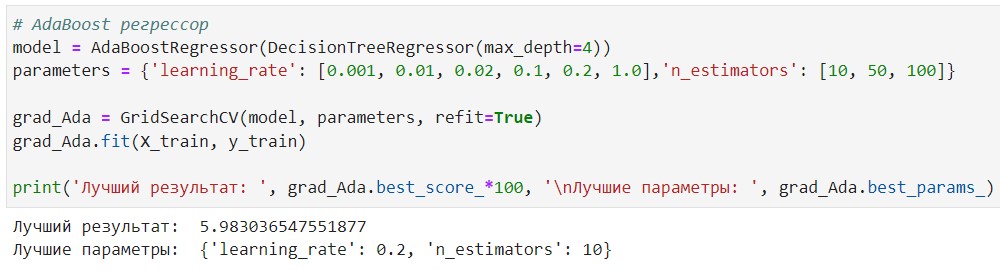
*Рисунок 25 – Предсказания на модели линейной регрессии*

В качестве метрик, определяющих качество работы модели в работе представлены средняя абсолютная ошибка MAE и коэффициент детерминации R2 .

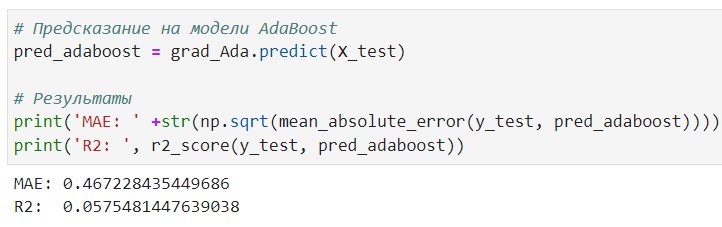
MAE показывает среднее значение ошибок и чем меньше это значение, тем лучше работает модель. R2 в свою очередь имеет обратный «эффект»: чем ближе значение коэффициента детерминации к 1, тем сильнее зависимость между признаками.

*Рисунок 26 – Сравнение рейтингов (оценок) предсказания модели*

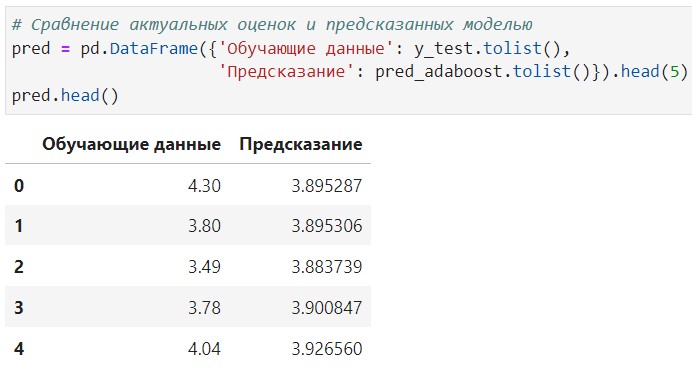
На рисунке 25 и 26 видно, что модель имеет большой процент ошибок и недостаточно точные предсказания.

Следующая модель представляет собой ансамблевый метод AdaBoostRegressor: DecisionTreeRegressor (рисунок 27).

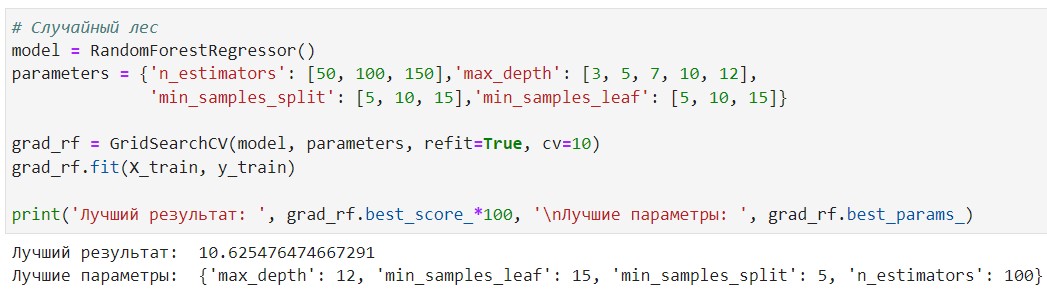
*Рисунок 27 – Модель AdaBoostRegressor (DecisionTreeRegressor)*

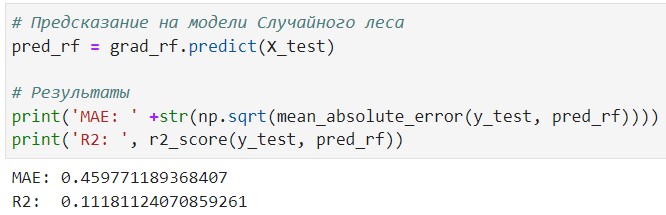
Ансамблевая модель, улучшающая работу регрессионного дерева решений была призвана улучшить результаты предыдущих попыток с обычным линейным регрессором.

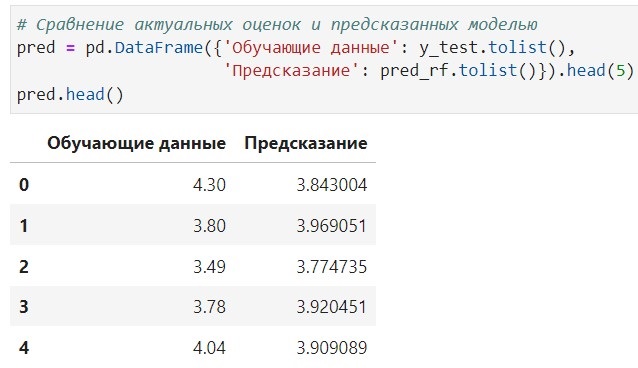
*Рисунок 28 - Предсказания на модели AdaBoostRegressor (DecisionTreeRegressor)*

Как видно на рисунках 28 и 29, результаты MAE и R2 улучшились, но незначительно.

*Рисунок 29 – Сравнение рейтингов (оценок) предсказания модели*

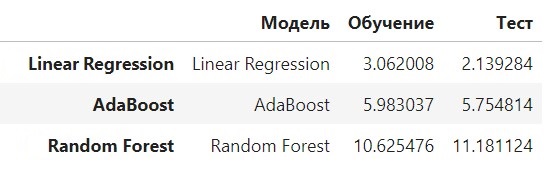
После обучения двух моделей складывается впечатление, что тенденции к улучшению результатов нет, поэтому можно экспериментировать дальше и применить еще один ансамблевый регрессионный метод RandomForest Regressor (рисунок 30).

*Рисунок 30 – Модель RandomForest Regressor*

*Рисунок 31 - Предсказания на модели RandomForest Regressor*

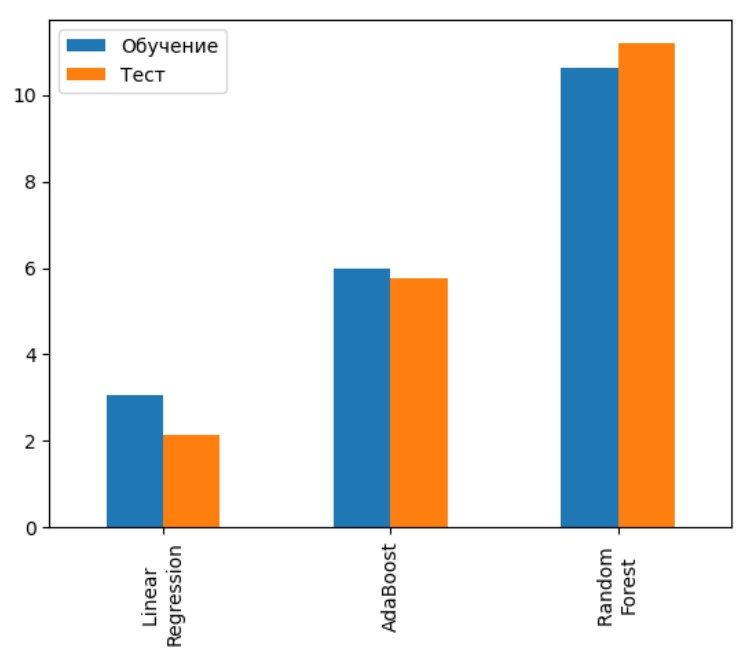
*Рисунок 32 – Сравнение рейтингов (оценок) предсказания модели*

Ансамблевая модель Случайного леса лишь немного улучшила показатель средней квадратичной ошибки, зато показала хороший подъем коэффициента корреляции.

После того, как обучение всех трех моделей завершено, настало время сравнить регрессионные модели между собой, чтобы наглядно увидеть, какой метод отработал лучше на предоставленном наборе данных.

*Рисунок 33 – Сравнение работы моделей в таблице*

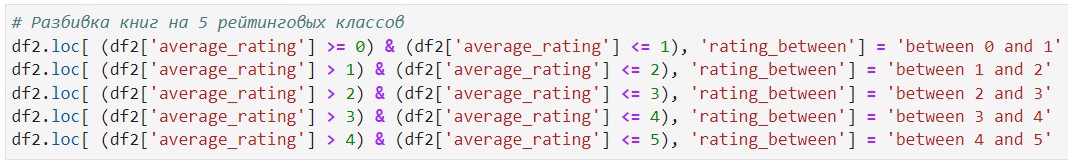
При построении таблицы использовались данные встроенных в инструмент кросс-валидации переменных: «best\_score» - «Обучение», «score» - «Тест», показатели которых равны коэффициенту детерминации R2.

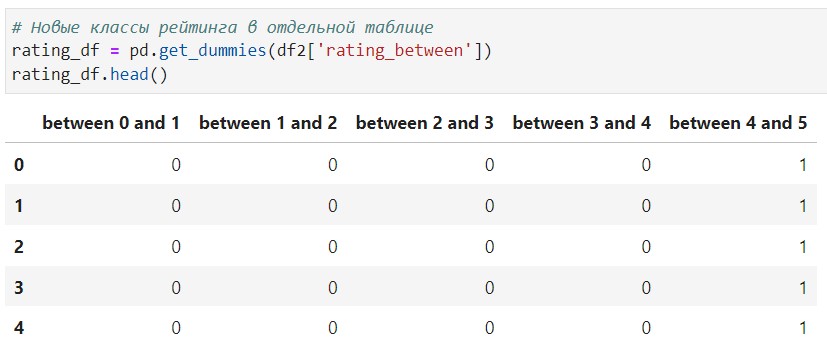
Для большей наглядности можно построить столбчатый график:

*Рисунок 34 – График сравнения данных работы моделей*

**2.3 Создание рекомендательной системы**

Для создания рекомендательной системы необходимо сделать вторую копию исходного набора данных, чтобы эксперименты никак не отразились на состоянии и содержании изначальной информации.

После этого происходит разделение оценок среднего рейтинга на пять классов (рисунки 35, 36).

*Рисунок 35 – Разделение среднего рейтинга на пять классов*

*Рисунок 36 – Новые классы рейтинга*

После работы с классификацией рейтингов необходимо привести в порядок колонку с категориальным признаком «Язык произведения», закодировать ее в числовые значения.

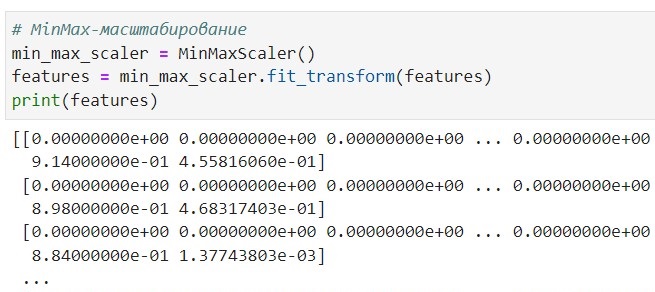
Кодирование колонки «Язык произведения» производится методом библиотеки Pandas – «get\_dummies», который создает новую таблицу атрибутов, присваивая каждому значению уникальный двоичный код (рисунок 37).

*Рисунок 37 – Новые признаки языков произведения*

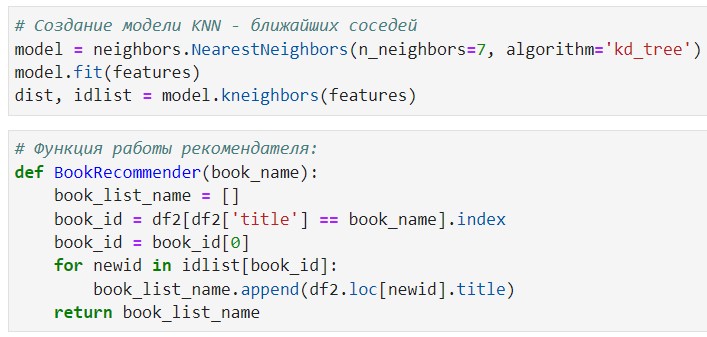
После того, как все данные преобразованы и закодированы, все необходимые для работы колонки объединяются в одну таблицу (рисунок 38):



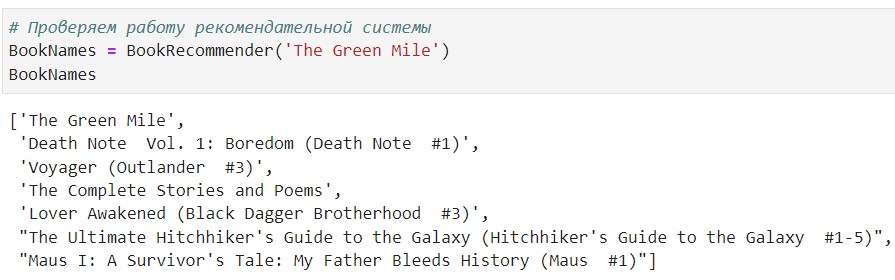
*Рисунок 38 – Данные для формирования рекомендательной системы*

 Перед созданием непосредственно самого рекомендатора данные нужно масштабировать. В данной работе это выполняется при помощи инструмента MinMaxScaler (рисунок 39).

*Рисунок 39 – Масштабирование данных*

Создание рекомендательной системы на предоставленном наборе данных будет представлять собой простой рекомендатор, основанный на средних оценках пользователей, выставленных для определенных книг (item-based). Алгоритм будет брать за основу показатель среднего рейтинга выбранной книги, затем на основе оценки этой книги предлагать пользователю произведения с похожим рейтингом.

*Рисунок 40 – Модель рекомендательной системы*

Для того, чтобы реализовать рекомендацию книг, ближайших выбранной по рейтингу используется алгоритм ближайших соседей KNN (рисунок 40).

*Рисунок 41 – Проверка модели рекомендательной системы*

Для проверки работы рекомендатора необходимо ввести имя книги, чтобы получить список, какие шесть книг находятся рядом искомой по среднему рейтингу (рисунок 41).

**Заключение**

Построенные регрессионные модели на основе данных набора Goodreads-books показали не слишком удовлетворительные результаты. Самым лучшим показателем средней ошибки MAE был ~0.4597 в модели Случайного леса (RandomForestRegressor). Это говорит о том, что модель делает примерно 46% ошибок в своих предсказаниях – такой регрессор выпускать «в поля» не стоит, с ним нужно еще поработать.

Значительную роль в формировании такого показателя средней ошибки внесла сама структура набора данных. Она изначально не предполагала создания регрессионных моделей. Датасет выложен в открытые источники с целью тренировки пользователей ресурса [www.kaggle.com](http://www.kaggle.com) в задачах классификации данных, возможно применении методов кластеризации и создания рейтинговых рекомендательных систем. Поэтому все созданные регрессионные модели, предсказывающие средний рейтинг книг по определенным показателям, получись не слишком точными.

Одновременно с этим попытка реализовать рекомендательную систему оказалось успешной: простой рекомендатель на основе item-based алгоритма, использующего средний рейтинг книг, подсказывает пользователю, какие книги есть с похожим рейтингом. Также удалось реализовать мини-программу по подбору книги в файле Jupiter Notebook.

По итогам работы можно сделать выводы о необходимости донастройки регрессионных моделей для получения более точных результатов предсказаний, хотя на первый взгляд кажется, что это невозможно. Но ведь именно попытки решить нерешаемое и есть основное кредо специалиста в области Data Science.

**Приложения**

1. Ссылка на открытый репозиторий с данной Выпускной квалификационной работой: <https://github.com/poluninamaria/DSprojects>

**Источники**

1. Платформа для соревнований специалистов в области анализа данных <https://www.kaggle.com> ;
2. Открытая интернет-энциклопедия: <https://www.wikipedia.org>
3. Интернет-ресурс для разбора и обсуждения IT-вопросов: <https://habr.com/ru/all/>
4. Шпаргалка по Python: <https://pythonru.com>
5. Описание работы библиотек Scikit-learn, официальный источник: <https://scikit-learn.org/stable/index.html>
6. Платформа для обмена файлами: <https://github.com>
7. <https://proproprogs.ru>
8. Проект студентов Яндекс-практикума, учебник по машинному обучению: <https://academy.yandex.ru/handbook/ml/>
9. Платформа с отзывами на книги (англоязычная): <https://www.goodreads.com>
10. «Простой Python. Современный стиль программирования», Билл Любанович, 2-е издание, O`Relly;
11. «Data science. Наука о данных с нуля», Джоэл Грасс, 2-е издание, O`Relly;