Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

Изображение выглядит как текст, керамические изделия, фарфор

Автоматически созданное описание

**Курсовая работа по дисциплине**

**«Методы машинного обучения»**

**на тему:**

**«Классификация книжных отзывов»**

**ИСПОЛНИТЕЛЬ:**

Фадеев Артем Александрович, ИУ5-31М

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

"\_\_"\_\_\_\_\_\_\_\_\_\_\_2021 г.

Москва 2021

# Оглавление

[Оглавление 2](#_Toc90950316)

[Задание 3](#_Toc90950317)

[Подготовка данных 4](#_Toc90950318)

[Загрузка датасета книг 4](#_Toc90950319)

[Устранение пропусков данных 4](#_Toc90950320)

[Обработка нестандартного признака 4](#_Toc90950321)

[Обработка категориальных признаков 5](#_Toc90950322)

[Нормализация числовых признаков 6](#_Toc90950323)

[Масштабирование признаков 6](#_Toc90950324)

[Отбор признаков 7](#_Toc90950325)

[Результат работы моделей 9](#_Toc90950326)

[AutoML 10](#_Toc90950327)

# Задание

1. Поиск и выбор набора данных для построения модели машинного обучения. На основе выбранного набора данных строится модель для задачи классификации.
2. Для выбранного датасета решить следующие задачи:
   1. устранение пропусков в данных;
   2. кодирование категориальных признаков;
   3. нормализацию числовых признаков;
   4. масштабирование признаков;
   5. обработку выбросов для числовых признаков;
   6. обработку нестандартных признаков (которые не является числовым или категориальным);
   7. отбор признаков, наиболее подходящих для построения модели;
3. Обучить модель и оценить метрики качества для двух выборок:
   1. исходная выборка, которая содержит только минимальную предобработку данных, необходимую для построения модели (например, кодирование категориальных признаков).
   2. улучшенная выборка, полученная в результате полной предобработки данных в пункте 2.
4. Построить модель с использованием произвольной библиотеки AutoML.
5. Сравнить метрики для трех полученных моделей.

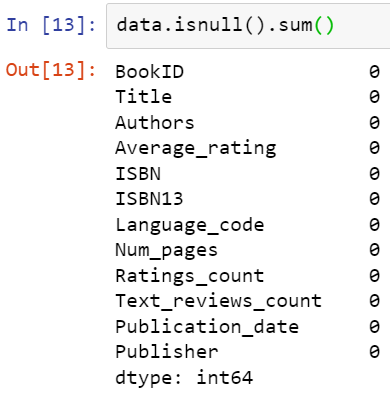
# Подготовка данных

## Загрузка датасета книг



## Устранение пропусков данных

Датасет без пустых значений признаков, поэтому данный этап его подготовки я пропустил. Хотя теоретически можно было бы заменить пропуски, например, медианными значениями.



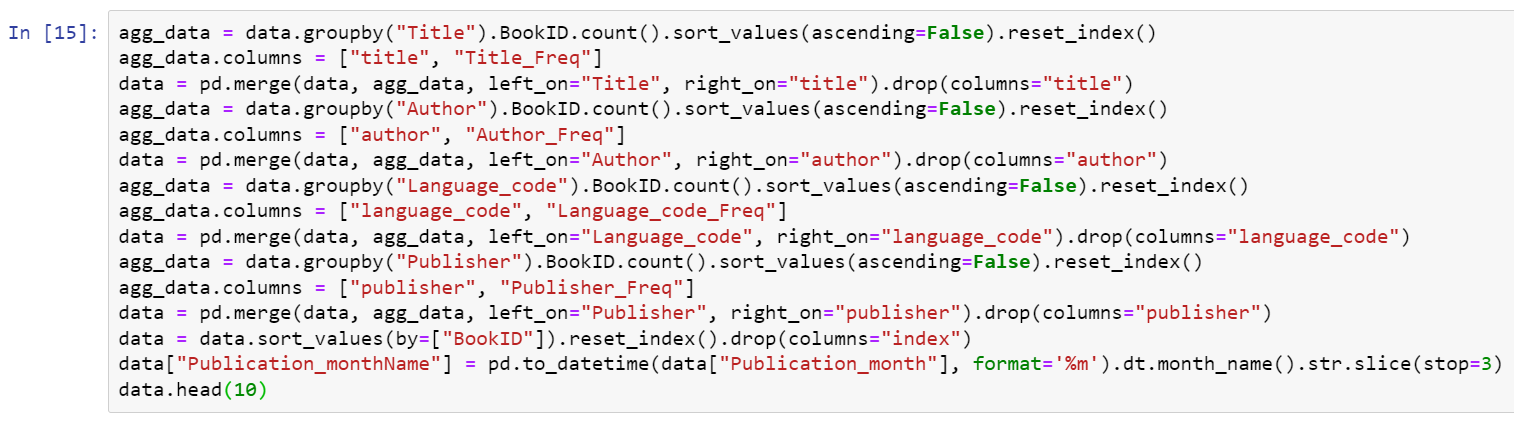
## Обработка нестандартного признака

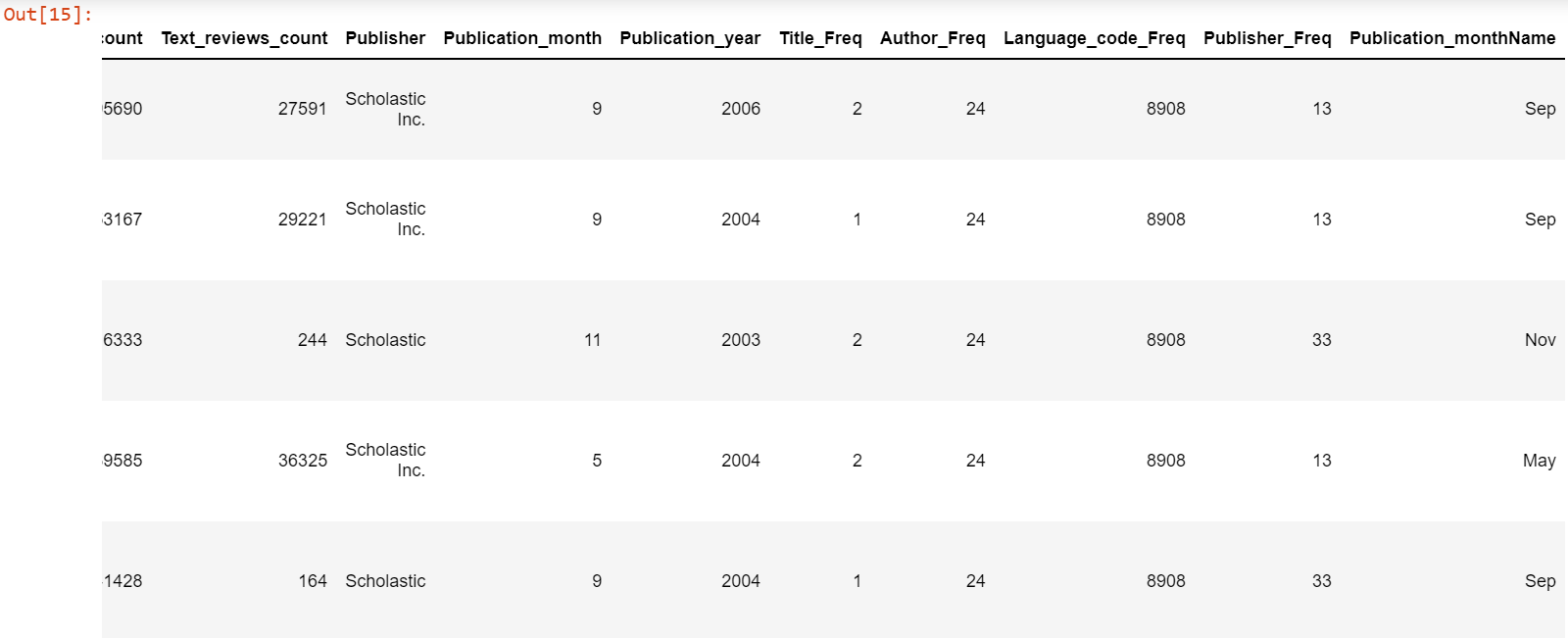
В данном блоке я переопределил столбец BookID, а также разделил дату публикации на месяц и год публикации, сделав из необычного признака два числовых. Из «месяцев» теперь легко выделить классы.



## Обработка категориальных признаков

Для данного этапа я не пользовался стандартными библиотеками, а создал собственный способ, который можно считать кодированием по частоте. Этот принцип, и его реализация в данной работе, предполагает замену какого-либо строкового значения количеством вхождений такого значения в данный признак (см. пару Publisher - Publisher\_Freq).

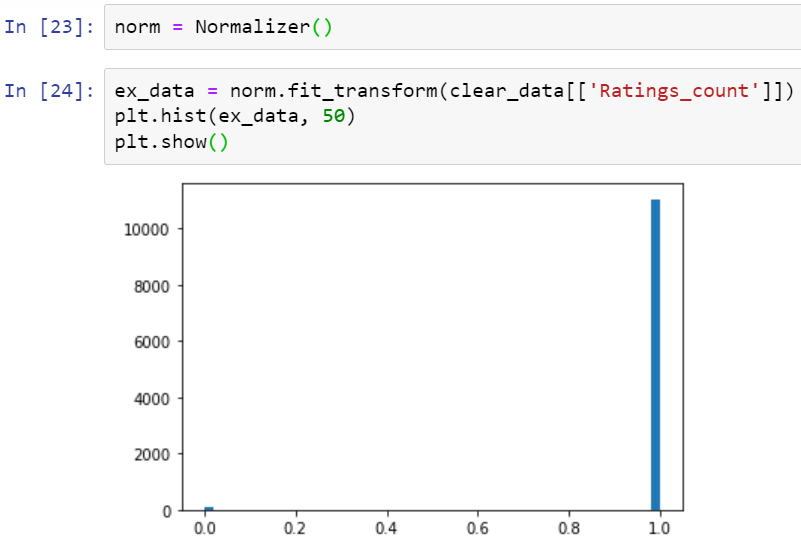




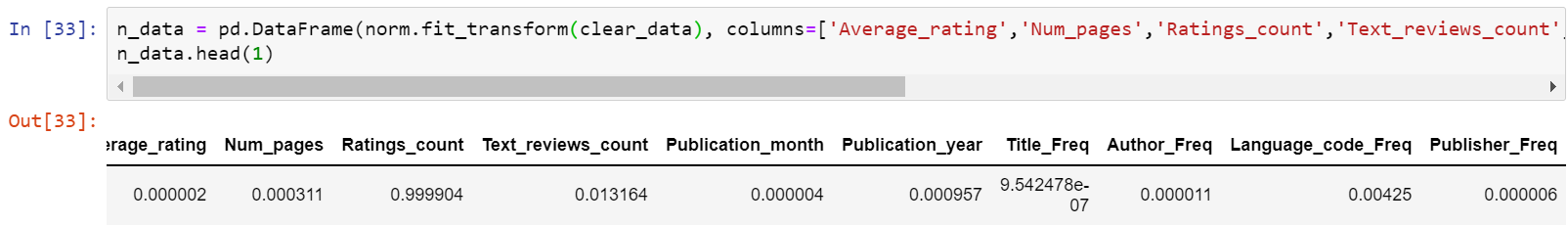
## Нормализация числовых признаков

Для этого использую стандартный метод, предварительно убрав лишние, уже обработанные признаки.



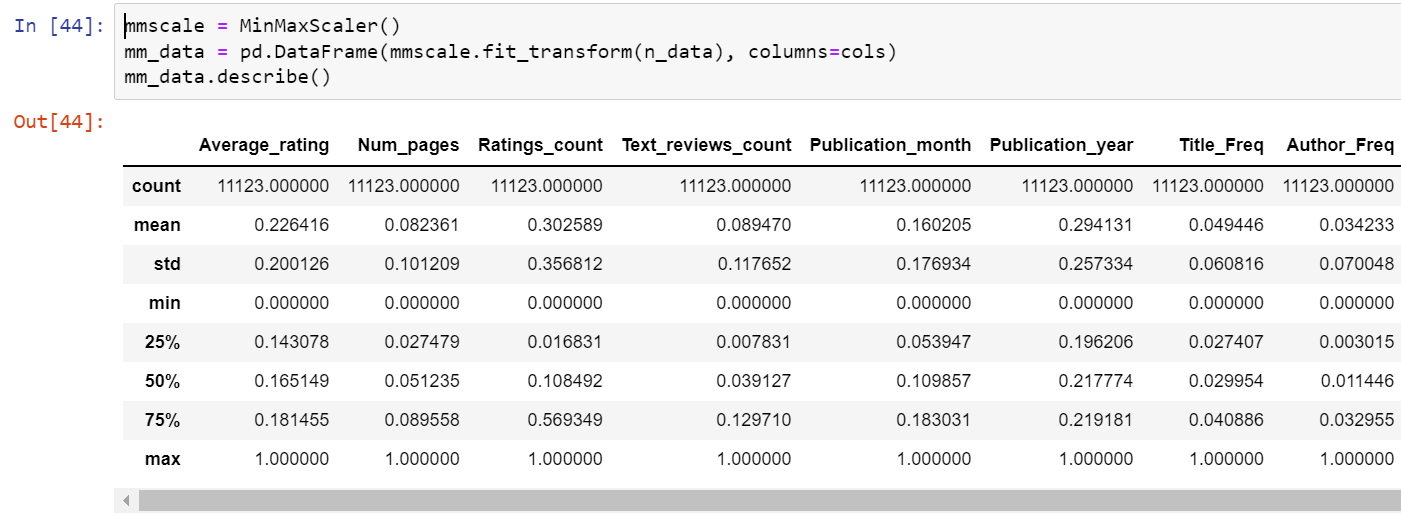


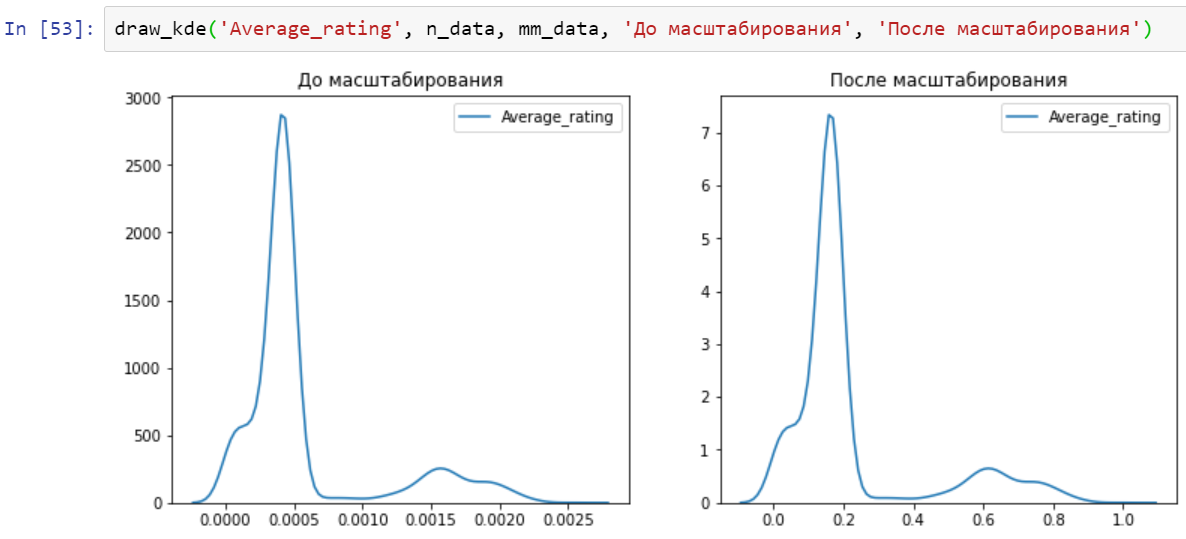
Аналогичную операцию произведу со всеми признаками.



## Масштабирование признаков

Для этого использую стандартный метод, предварительно убрав лишние, уже обработанные признаки.

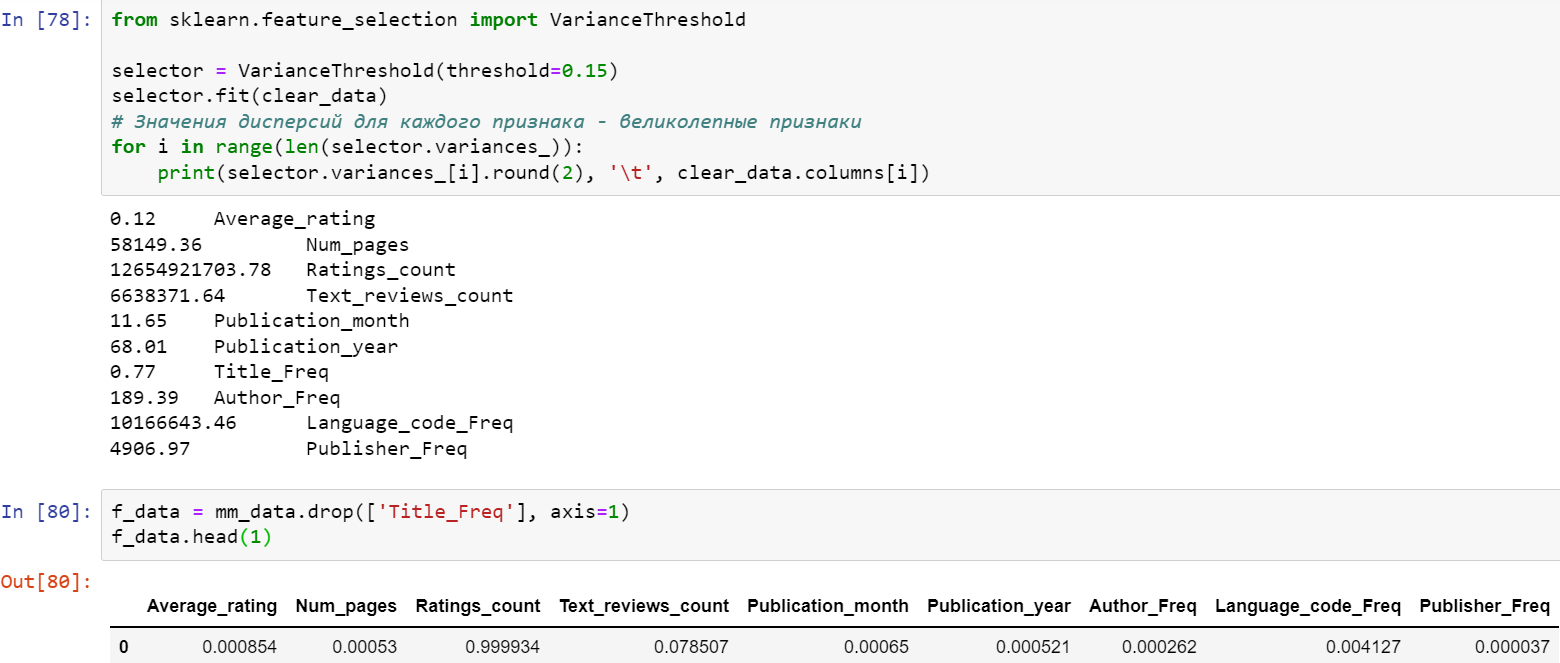




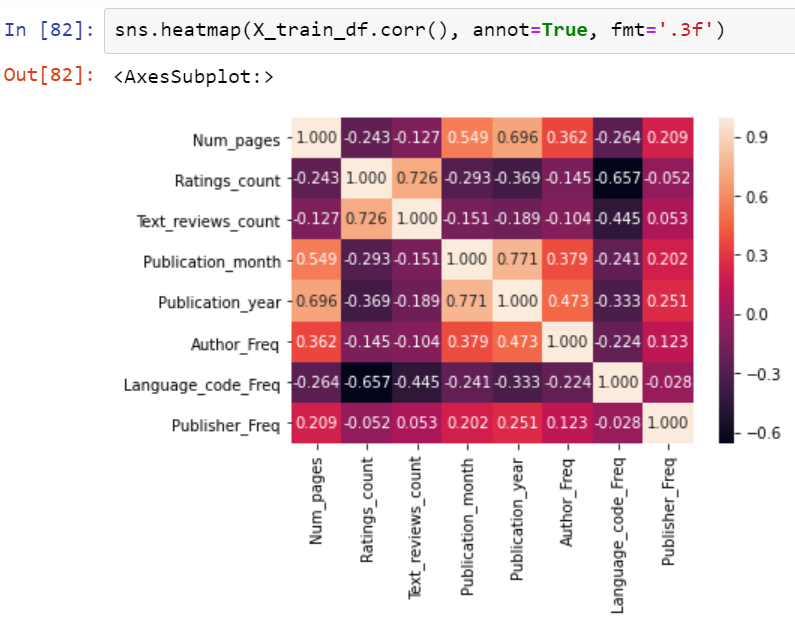
## Отбор признаков

Для этого использую метод фильтрации, основанный на удалении константных или почти константных признаков.

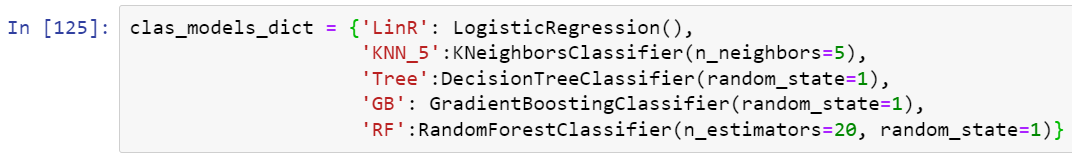
По итогу можно избавиться от двух колонок: Заглавие книги. Средний рейтинг является будущим целевым признаком, поэтому его не удаляю.

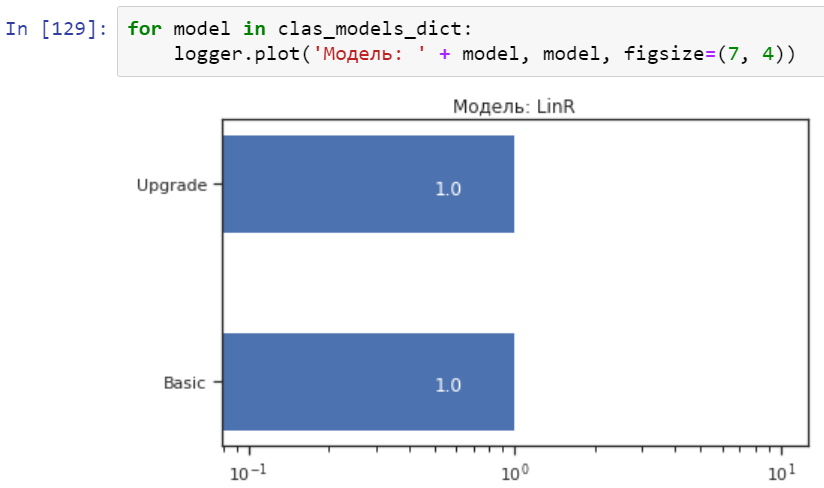


Также для фильтрации можно применить отсев признаков, состоящих в какой-либо группе коррелирующих признаков (например, у которых корреляция между друг другом > 0,9). Однако, в данном наборе данных таких признаков нет, поэтому не можем применять этот метод.



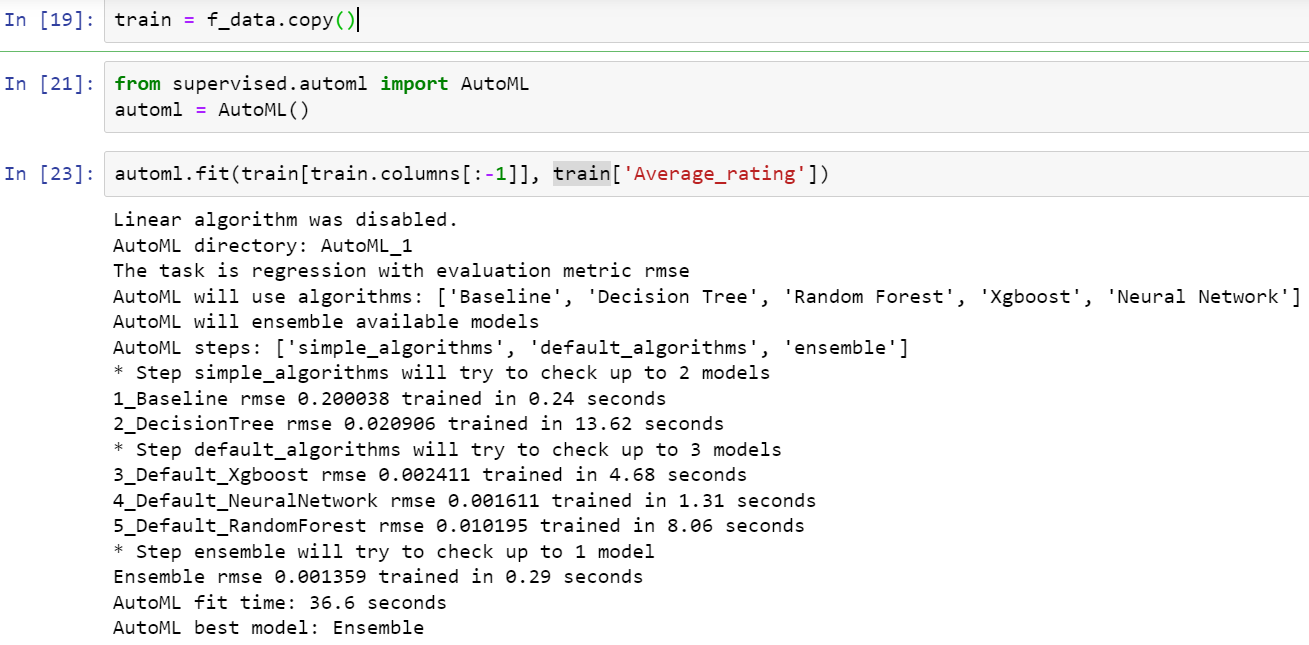
# Результат работы моделей





Для KNN\_5 - 0.94. Для остальных - 1.

# AutoML



Для полностью подготовленного датасета без лишнего признака «Заглавие книги» точность Ансамблевой модели (обучилась за 0.29 секунд) составляет 99,8641% при обучении самой AutoML в 36.6 секунд - неплохо!

Если этот признак не удалять, то обучение Ансамблевой модели будет происходить быстрее (0.21 секунды), точность - 99,8228% (разница более 0,04 процента), а обучение самой AutoML в 27.98 секунд - тоже неплохо.

