Университет ИТМО

Практическая работа №5-6

по дисциплине «Визуализация и моделирование»

Автор: Голуб А. Л.
Поток: ВИМ 1.2
Группа: К3243
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

Датасет: Goodreads-books

1 Машинное обучение

Задача: обучить модель линейной регрессии, чтобы пресказать значения столбца **average_rating** - средний рейтинг книги на сайте.
Предварительно была проведена предобработка:

- Исправлены некорректные данные в столбце с датой публикации.
- Значения в этом столбце заменены на число дней, прошедшее с публикации самой ранней книги в датасете.
- Отдельно выделен столбец с годом публикации.
- Коды языков публикации и издательства закодированы числами.
- Добавлен столбец с числом авторов.
- Удалены столбцы, не несущие информацию, которая не будет использоваться для обучения модели (название книги, код ISBN и др.)
- Нормализованы данные в столбцах с независимыми переменными.

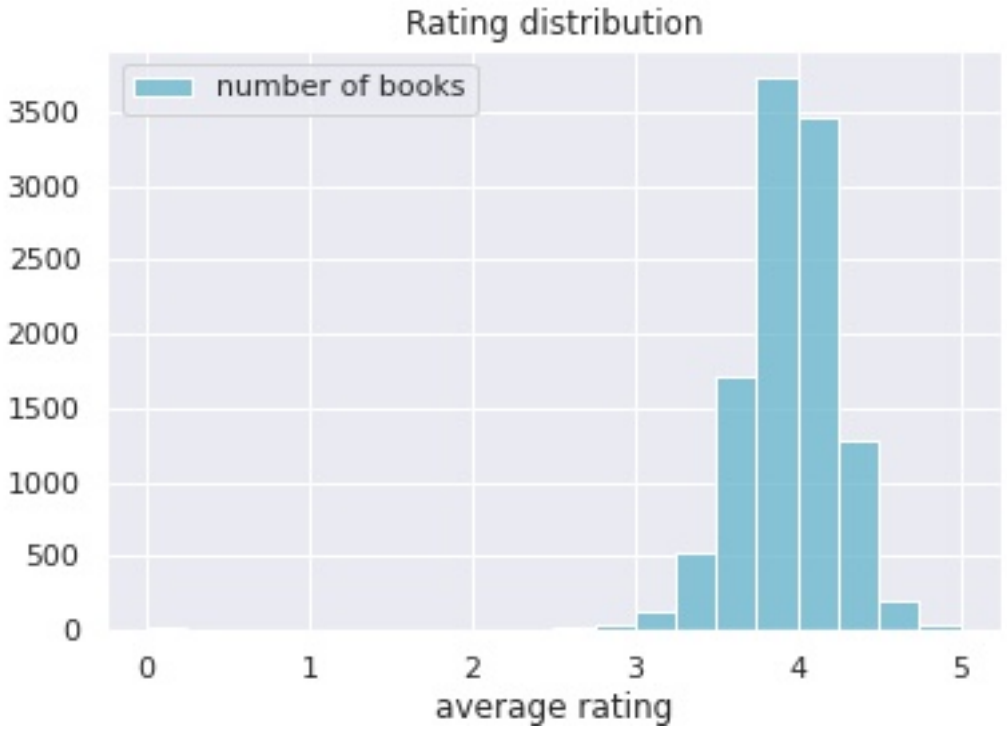
Для машинного обучения использовалась модель **LinearRegression** из sklearn. Значения коэффициента детерминации получились низкими - 0.029 на тренировочной выборке и 0.19 на тестовой. Следовательно, данные **крайне плохо** описываются моделью линейной регрессии.

Затем была обучена еще одна модель линейной регрессии с целью предсказать **year** - вероятный год издания книги, - считая известным ее средний рейтинг. (В датасете представлены 87 различных значений года в промежутке 1900 - 2020.) Такая модель обладает гораздо более высоким коэффициентом детерминации - 0.99 на тренировочных и тестовых данных, поэтому с помощью этой модели можно достаточно точно восстановить год издания книги, обладая другими сведениями о ней.

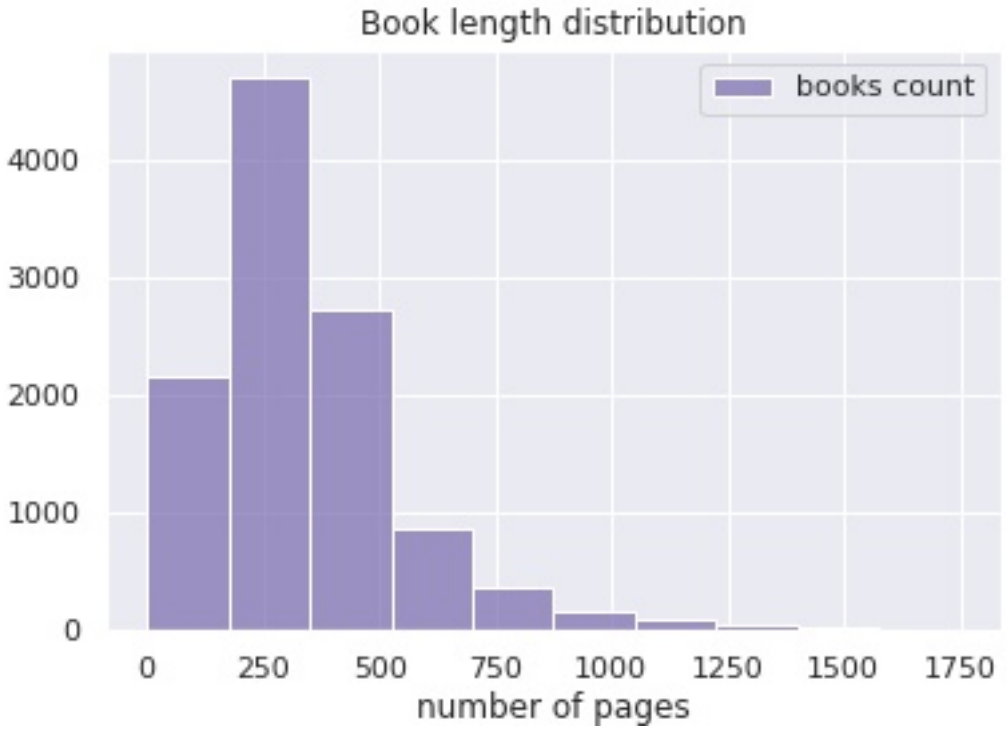
2 Data Storytelling

Задача: создать четыре-пять визуализаций, наиболее полно отражающих закономерности в данных, учитывая знания о них, полученнные на всех этапах работы.

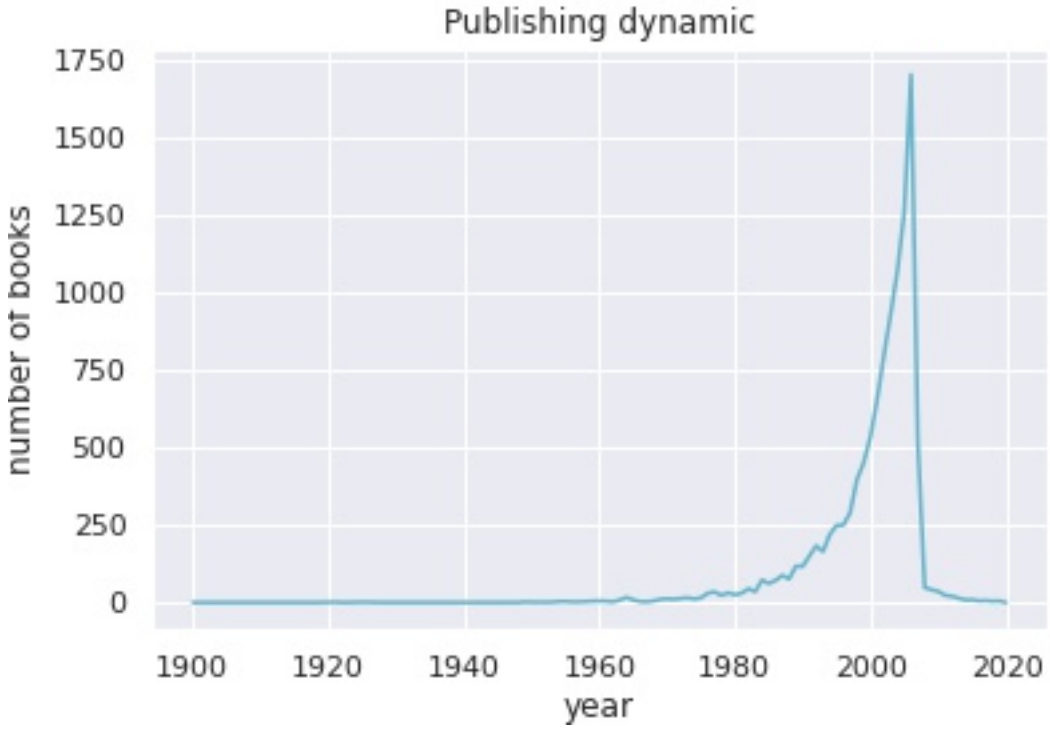
Распределение книг по среднему рейтингу.



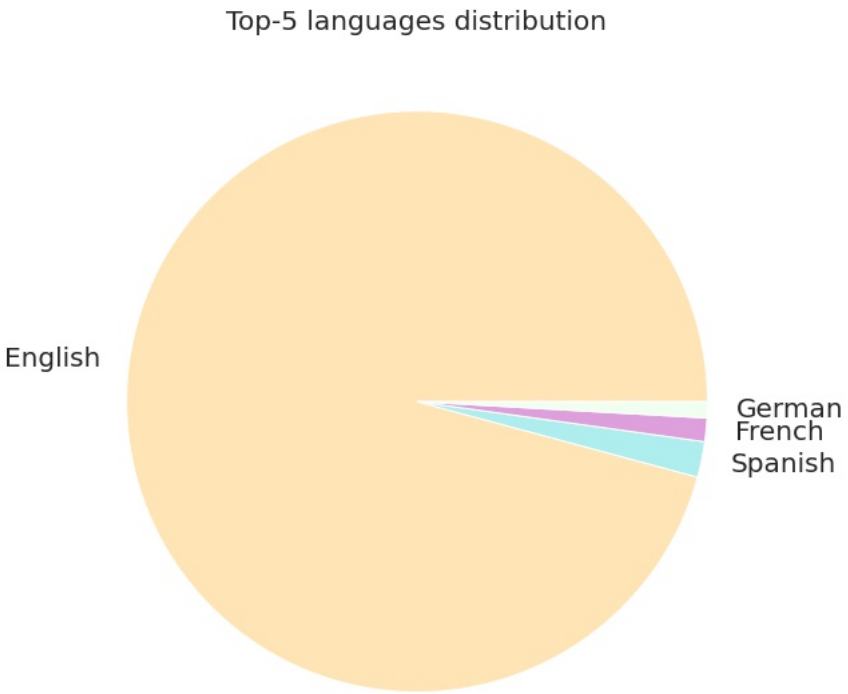
Распределение книг по числу страниц.



Динамика публикаций во времени.



Самые популярные языки книг.



Jypyter notebook