Университет ИТМО

Практическая работа №6

По дисциплине “Визуализация и моделирование”

Автор: Лайок Олег Владимирович

Поток: 1_2

Группа: К3242

Факультет: ИКТ

Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

Для решения задач машинного обучения я решил взять те же данные, которые я использовал до этого. Чтобы доказать, что данные подходят для этого проанализируем их по схеме CRISP-DM

1. Бизнес-анализ:
   Цель проекта: создать модель для предсказания цены, рейтинга и количества отзывов для конкретного жанра книги, чтобы улучшить продажи книг таких жанров (зная примерно по какой цене покупают книги конкретных жанров и какая аудитория у книг, рассчитывая из количества отзывов и рейтинга, будет легче подготовить книгу к выходу)
   Риски: небольшой объем данных, есть риск составить некорректную или не точную модель. Для оценки модели будем использовать метрику
2. Анализ данных: тип - 3rd party data (сайт amazon.com)
   Описание датасета: Топ 50 книг бестселлеров на сайте Amazon, в период с 2009 по 2019 год, с краткой характеристикой каждой книги в топе.
   Содержание датасета:
   Name - название книги, строка, в датафрейме данное поле имеет тип object что является верным, интервала определеных значений не имеет

   Author - имя автора, строка, в датафрейме данное поле имеет тип object что является верным, интервала определеных значений не имеет

   User Rating - средний рейтинг пользователей, число с плавающей точкой, в датафрейме данное поле имеет тип float64 что является верным, интервал значений (0.0,5.0]

   Reviews - число отзывов на книгу, целое число, в датафрейме данное поле имеет тип int64 что является верным, заданного интервала значений не имеет

   Price - цена книги, целое число, в датафрейме данное поле имеет тип int64 что является верным, интервал значений >0

   Year - год в котором книга попала в топ, целое число, в датафрейме данное поле имеет тип int64 что является верным, диапазон значений {2009,2010,2011,2012,2013,2014,2015,2016,2017,2018,2019}

Genre - жанр книги, строка, в датафрейме данное поле имеет тип object что является верным, может принимать только значения Fiction/Non Fiction

Subgenre - поджанр книги, строка, в датасете имеет тип object что является верным, диапазон значений {'Thriller', 'Comedy', 'Non Fiction', 'Comics', 'Games', 'Young Adult Fiction', 'Historical', 'Children', 'Science fiction', 'Novel', 'Fantasy', 'Horror', 'Drama'}

Название колонки в датасете	Данные	Тип данных
Name	Название книги	Текстовые
Author	Имя автора	Текстовые
User Rating	Рейтинг пользователей	Числовые
Reviews	Количество отзывов	Числовые
Price	Цена	Числовые
Year	Год, когда книга в топе	Числовые
Genre	Жанр (только два – fiction/non-fiction)	Текстовые
Subgenre	Поджанр	Текстовые

3. Подготовка данных:
   После анализа данных (в работе 3) было выявлено, что пустых значений нет, артефактов нет, все значения в пределах допустимых значений и имеют корректный тип.
   Отбор данных: самой главной связью для бизнеса будет связь поджанров и цены книги данного поджанра, соответственно нас будут интересовать столбцы Subgenre и Price. Так же, для предсказания количества людей купивших книгу данного жанра можно исследовать связь между поджанром книги и количеством отзывов и рейтингом у книг данного поджанра, для этого будем анализировать столбцы Reviews и User Rating

4. Моделирование
   Перед нами стоит задача регрессии: значения столбцов Price, Rating и Reviews нужно научится предсказывать для соответсвующих значений из столбца Subgenre. Т.к. книг в каждом поджанре несколько, модель по данной тестовой выборке сможет предсказывать наиболее вероятное значение цены, рейтинга и количества отзывов для каждого поджанра.
   Выбор модели регрессии:
   Т.к. данных не так много (всего 550 строк) нужно опытным путем проверить, какая модель будет лучше предсказывать значения. Было решено применить модели линейной регрессии и random forest
   Ссылка на код с комментариями по применению моделей и оценки их результатов:
   https://colab.research.google.com/drive/1lWYys17yaBwKoqG_6Q-4PtqLPMEcTIRv?usp=sharing
   Ссылка на красивую визуализацию результатов регрессии:
   https://app.flourish.studio/story/892560/edit
5. Оценка результатов:
   Используя наш датасет можно с довольно высокой точность предсказать цену и примерный рейтинг для каждого поджанра, однако число рецензий с таким количеством данных предсказать не удастся. Лучше всего с предсказанием результатов справляется модель Случайного Леса.
6. Внедрение:
   Полученная модель по предсказанию цены и рейтинга книги по ее поджанру может быть использована издателями и маркетплейсами при выставлении книг на продажу. Однако, данную модель следует в будущем дообучать на больших объемах данных чтобы повысить его точность.