Курсовая работа на тему:

Использование моделей регрессии для прогнозирования цен на произведения искусства

декабрь 2019

Студент 231 группы:
Мироненко Ф.Д.

Научный руководитель:
Григорьев Д.А.

Санкт-Петербургский Государственный Университет

1 Введение

В нашей работе нам удалось присоединиться к международной команде, занимающейся исследованиями в области анализа рынка художкственных произведений. Основоной целью исследования является создание модели, определяющей возможную цену картины на основе различных признаков: тональности, изображённых объектов, года написания, бывших владельцев и т.д.

Главной задачей на начальном этапе был сбор датасета. Для этих целей использовался язык Python3.7 и среда разработки Jupyter Notebook. С исходным кодом, а также с полученными датасетами можно ознакомиться по ссылке:

https://github.com/FomaMironenko/Projects/tree/master/Project_sem3

В дальнейшем названия всех файлов выделены курсивом и приводятся в соответствии с их именами на GitHub.

2 Этапы работы

Изначально для работы нам был предоставлен файл *Source.txt*, содержащий информацию (автор, название, дата последней продажи, цена) для 1000 произведений искусства. Первым делом он был приведён к более удобному для обработки виду в *arts.csv*. Далее работа разделилась на две части по сбору данных по всем художникам и по всем картинам, доступным из списка.

2.1 Данные по художникам

Для получения информации по художникам была использована платформа mutualart.com [1]. Сперва из *arts.csv* были выделены 161 различных художников, и каждому из них при помощи скрипта *get_artists_url.ipynb* был сопоставлен адрес соотвестствующей страницы с mutualart.com [1]. Если корректный адрес не был найден, соотвествующее поле заполнялось пустой строкой. После этого для каждого художника с непустой ссылкой при помо-

115	Cy Twombly	https://www.mutualart.com/Artist/Cy-Twombly/C5...
116	Lucio Fontana	https://www.mutualart.com/Artist/Lucio-Fontana...
117	Paul Gauguin	https://www.mutualart.com/Artist/Paul-Gauguin/...

Рис. 1: artists.csv после применения get_artists_url.ipynb

щи скрипта *get_artist_info.ipynb* с соответсвующего адреса были получены поля `Country, Born, Died, Info` и записаны в таблицу *artists.csv*.

115	Cy Twombly	https://www.mutualart.com/Artist/Cy-Twombly/C5...	American	1928	2011	[\r\n Edwin Par...
116	Lucio Fontana	https://www.mutualart.com/Artist/Lucio-Fontana...	Italian	1899	1968	[\r\n Influenti...
117	Paul Gauguin	https://www.mutualart.com/Artist/Paul-Gauguin/...	French	1848	1903	[\r\n French ar...

Рис. 2: artists.csv после применения get_artist_info.ipynb

2.2 Данные по картинам

Источником информации для датасета по картинам послужил сайт аукционного дома Christies [3]. Сбор данных осуществлялся при помощи кода из файла *get_ art_ info.ipynb*. Для каждой строки из *arts.csv* просматривалась поисковая выдача сайта для соответствующего произведения искусства, и из первых 30ти вариантов выбирался наиболее релевантный. При этом, как видно из Рис. 3, результаты поиска не всегда содержали искомый объект. За эту чать программы отвечает функция `get_correct_url`, принимающая на вход имя художника и название картины. Для выбора наиболее релевантного результата имя художника и название картины из датасета последовательно сравнивались с соответствующими полями из вариантов выдачи при помощи метода `SequenceMatcher.ratio()` из библиотеки `difflib`, и выбирался

$$argmax\{ratio(x.p, dataset\_p) : x \in search\_result \ \& \ ratio(x.a, dataset\_a) > 0.9\}$$

где p - название картины, a - имя художника. `get_correct_url` возвращает строку с адресом подходящей страницы, которая поступает на вход в функцию `parce`, добывающую значения для полей `Estimate, Description, Provenance`. Обработанные таким образом данные записывались в файл *artsnew.csv*.

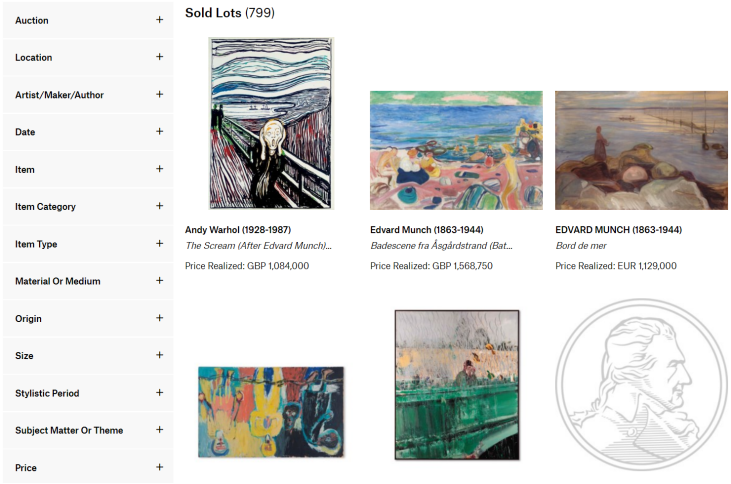

Рис. 3: Поисковая выдача для запроса Edvard Munch The Scream

3 Результаты

Таким образом, получены два файла: *artsnew.csv*, содержащий таблицу 1000×9, и *artists.csv* с таблицей 161×6, которые будут использованы в дальнейшем для обучения модели.

4 Использованное ПО

Пакеты для `Python`:

- `numpy 1.18.0`
- `pandas 0.25.3`
- `requests 2.22.0`
- `bs4 0.0.1`
- `fake_useragent 0.1.11`
- `difflib`

Использованные ресурсы

[1] https://www.mutualart.com/

[2] https://www.sothebys.com/en/

[3] https://www.christies.com/