Data Science Final Project Nhóm 15

Thành viên nhóm 15

- Nguyễn Thanh Tuấn 1612744
- Hoàng Xuân Trường 1612899

Thông tin chung

Tệp

- **Slide** :https://docs.google.com/presentation/d/1Nq5LlfNVe-uooOjxUx5RLwBN-Aha2yDpy-C7NqmzX9Y/edit?usp=sharing

Ý tưởng

- Dự đoán giá laptop, tablet từ các thông tin cấu thành (hãng, CPU, ram, gpu,...)
    - Input: các thông tin của mấy tính: hãng, CPU, ram, ...
    - Output: Giá sản phẩm.

Nguồn cảm hứng và ứng dụng

- Người dùng: Gợi ý giá sản phẩm cho người dùng khi người dùng muốn mua 1 sản phẩm laptop, tablet với cấu hình mong muốn.
- *Model dự đoán giá này có gì đặc biệt:*

> Model sẽ giúp người dùng tự build sản phẩm từ những linh kiện khác nhau nên người dùng không cần phải tra cứu từng linh kiện để lắp ráp. Người dùng có thể tự tìm kiếm các link kiện nhưng đó là với người dùng có kiến thức nhất định về các cấu hình máy (loại nào, hãng nào, thế hệ nào,..). Vì vậy, mô hình sẽ tốt cho *(a)* tiết kiệm thời gian tìm kiếm từng thành phần cấu hình laptop một cách chi tiết nhưng vẫn có thể có chi phí tối ưu và *(b)* dễ sử dụng với những người dùng có không có kiến thức chi tiết về cấu hình laptop đã có sẵn trong bộ dữ liệu

Dữ liệu

- Lấy từ 2 nguồn: Bestbuy và Amazon
- API: Không lấy được đủ thông tin -> request_html và selenium
- Check tính hợp lệ của dữ liệu:
    - https://www.amazon.com/robots.txt
    - https://www.bestbuy.com/robots.txt
- Dữ liệu hiện tại:
    - Bestbuy: hơn 1000 items
    - Amazon: hơn 5000 items
- Vấn đề hiện tại:
    - Chọn thuộc tính phù hợp
    - Thuộc tính sản phẩm 2 nguồn khác nhau, format khác nhau, cần đồng bộ (khó ghép lại) -> giải quyết: ????
    - Dữ liệu rác: Thuộc tính không chuẩn, thiếu

Thông tin dữ liệu

Một số thông tin cơ bản được crawl:

- Price: giá của laptop
- Screen size: kích thước của màn hình
- RAM: bộ nhớ RAM
- Brand Name: hãng laptop
- Item Weigth: Khối lượng
- Operating System: Hệ điều hành của máy
- Color: màu của sản phẩm
- Processor Brand: hãng sản xuất CPU (Intel, AMD, Mediatek,..)
- Processor Count: số lượng nhân
- Computer Memory Type: Loại RAM (DDR3, DDR4,...)
- ... ~ Tất cả: 168 thông tin (Đối với dữ liệu của BestBuy).

Preprocessing

Dữ liệu được chọn là **BestBuy** (bỏ qua **Amazon** để tránh nhiễu).

Chọn một số thuộc tính được để tiền xử lý. Các thuộc tính khác do tỷ lệ thiếu quá cao nên sẽ bị bỏ qua. Các thuộc tính tiêu biểu:

- Giá
- Kích thước màn hình
- Độ phân giải

- Ram
- Chip xử lý
- Bộ nhớ
- Kích thước máy
- Khối lượng
- ...

~ 26 thuộc tính

Thuộc tính thiếu

Các thuộc tính thiếu được thay thế bởi các giá trị trung bình (cho numeric), giá trị có tầng suất xuất hiện nhiều nhất (cho object/string) tùy vào trường nào đang xét sẽ có các giá trị được chọn phù hợp nhất với thuộc tính đó.

Dataset

- Dữ liệu chưa được clean và preproccessing (file `*.csv`) nằm trong thư mục `/dataset/Amazon` cho dữ liệu của Amazon và `/dataset/BestBuy` cho dữ liệu của trang web Bestbuy.
- Dữ liệu đã qua xử lý nằm ở thư mục `preprocessing`, script xử lý là các file `preprocessing.ipynb`
- Nhóm thực hiện chia dữ liệu thành 2 tập chính train (90%) và test (10%).

Các thuộc tính được rút trích:

```
Index(['Price', 'Key_Specs__Screen_Size', 'Key_Specs__Touch_Screen',
       'Key_Specs__Storage_Type', 'Key_Specs__System_Memory', 'RAM_type',
       'RAM_speed', 'Processor', 'Graphics__Graphics', 'Screen_resolution1',
       'Screen_resolution2', 'Key_Specs__Operating_System',
       'Key_Specs__Battery_Type', 'General__Color_Category', 'General__Brand',
       'Feature__Keyboard_Touch_Screen', 'Feature__Backlit_Keyboard',
       'Feature__Mac_Features', 'Port_Number_USB_Ports',
       'Display__Display_Type', 'Storage__eMMC_Capacity',
       'Storage__Solid_State_Drive_Capacity', 'Dimension__Product_Depth',
       'Dimension__Product_Height', 'Dimension__Product_Weight',
       'Dimension__Product_Width'],
      dtype='object')
```

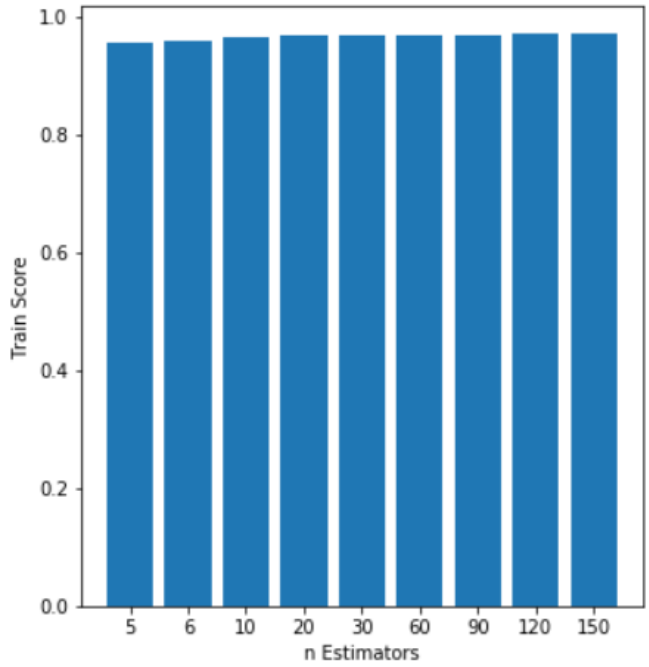
Một mẫu đại diện:

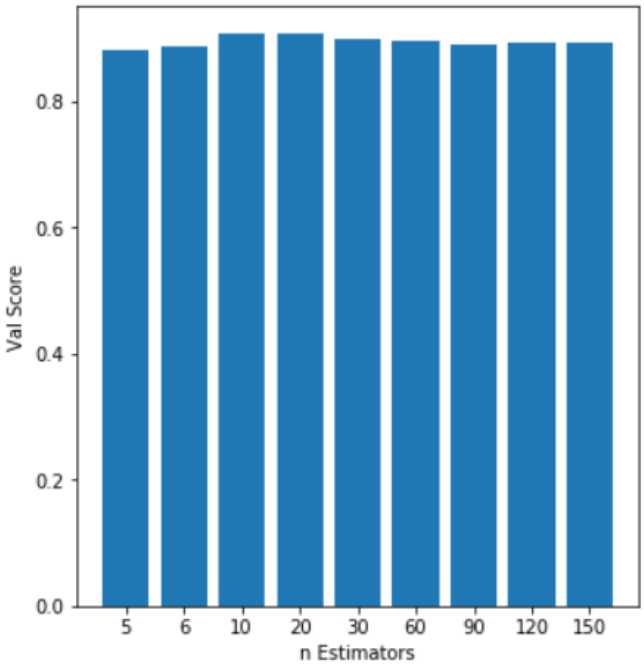
```
Price                                                    1299.99
Key_Specs__Screen_Size                                      13.3
Key_Specs__Touch_Screen                                       no
Key_Specs__Storage_Type                                      ssd
Key_Specs__System_Memory                                       8
RAM_type                                                  lpddr3
RAM_speed                                                   2133
Processor                                 intel 8th generation core i5
Graphics__Graphics                        intel iris plus graphics 645
Screen_resolution1                                          2560
Screen_resolution2                                          1600
Key_Specs__Operating_System                                mac os
Key_Specs__Battery_Type                               lithium-polymer
General__Color_Category                                space gray
General__Brand                                              apple
Feature__Keyboard_Touch_Screen                        apple touch bar
Feature__Backlit_Keyboard                                     yes
Feature__Mac_Features          force touch trackpad, siri, touch id sensor
Port_Number_USB_Ports                                         NaN
Display__Display_Type                                         lcd
Storage__eMMC_Capacity                                          0
Storage__Solid_State_Drive_Capacity                           256
Dimension__Product_Depth                                     8.36
Dimension__Product_Height                                    0.59
Dimension__Product_Weight                                    3.02
Dimension__Product_Width                                       12
Name: 10, dtype: object
```

Model

- Tập dữ liệu train được chia thành 2 tập chính: phần train (90% tập train, phục vụ cho quá trình train mô hình), phần validation (10% tập train, phục vụ đánh giá mô hình trong quá trình train).
- Nhóm áp dụng một vài thuật toán trên tập sữ liệu đang có. Kết quả đạt tốt nhất với thuật toán Randomforest:
- Thử các trường hợp n_estimators khác nhau, n_estimators=10 (và 20) cho kết quả tương đối tốt. Tuy nhiên, nhóm chọn n_estimators=10 cho mô hình cuối cùng.
- Độ chính xác trên tập train khoảng 96%, tập validation khoảng 90%.

Chọn n_estimators phù hợp cho dữ liệu:





Testing

- Nhóm thực hiện test trên 10% dữ liệu đã chia từ trước.
- Model chọn: Randomforest - n_estimator=10.
- Độ chính xác trên tập test: ~85%.
- Một số đánh giá: - Dữ liệu hơi ít, nên mô hình thực sự đủ để predict giá gần với giá gốc. - Các items có giá càng cao, thì sự chênh lệch giá predict và giá thực tế càng cao (Một phần do số lượng các items có giá cao trong tổng dữ liệu khá ít).

So sánh kết quả predict và kết quả thực:

	predicted	true	Difference
13	1071.88	1079.99	-8.11
54	705.94	699.99	5.95
63	1242.28	1269.99	-27.71
35	1507.87	1499.99	7.88
53	1096.86	769.99	326.87
16	573.01	499.99	73.02
25	1532.97	1699.99	-167.02
0	287.78	328.99	-41.21
14	583.97	599.99	-16.02
76	1305.52	1399.99	-94.47

Cách làm việc nhóm

Giai đoạn 1

Thời gian 24/12/2020

Địa điểm: Facebook

Task	Member
Lựa chọn chủ đề, tìm nguồn	Hoàng Xuân Trường, Nguyễn Thanh Tuấn
Get links	Hoàng Xuân Trường
Crawl dataset	Hoàng Xuân Trường
Fix Crawl links	Nguyễn Thanh Tuấn

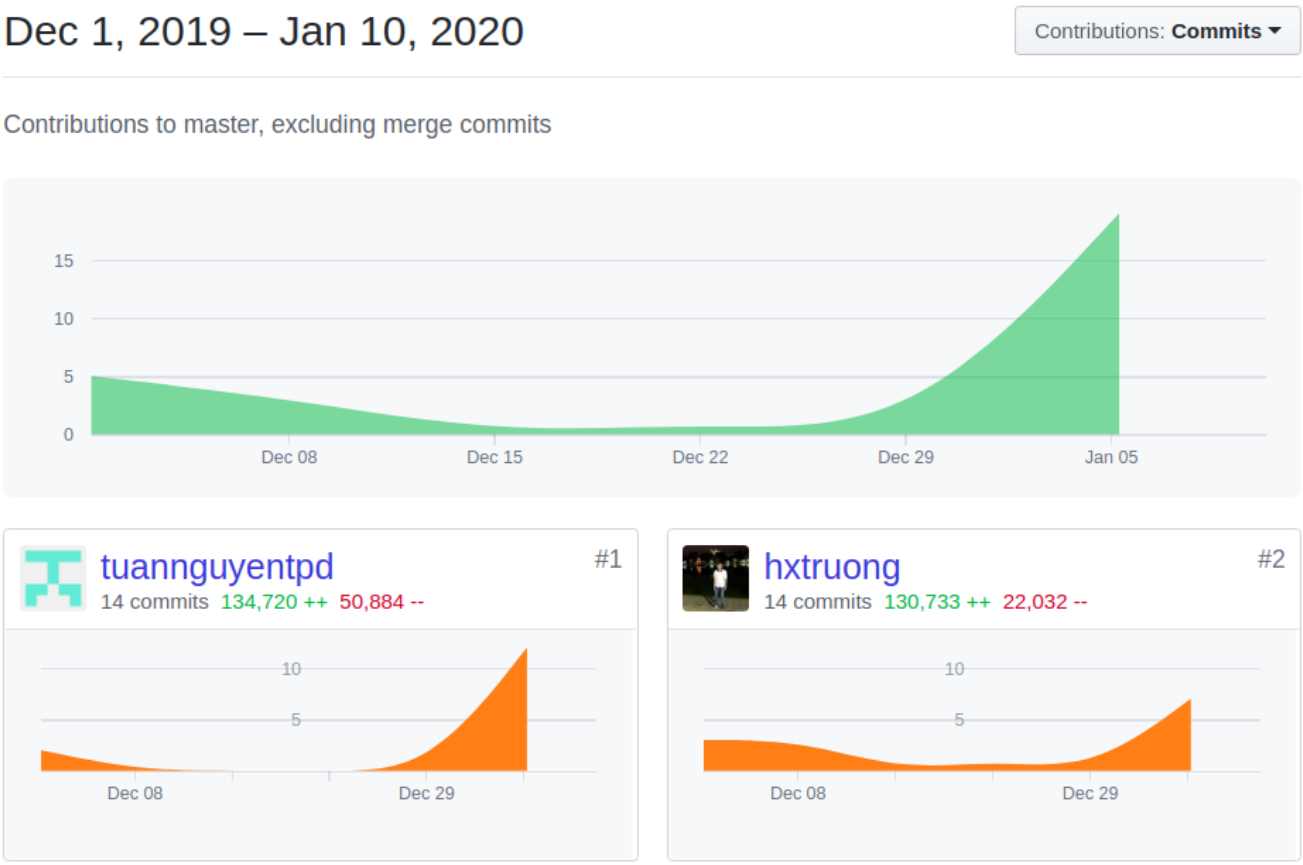
Gian đoạn 2

Thời gian: 2/1/2020

Địa điểm: Facebook, sảnh nhà I

Task	Member
Extract feature	Hoàng Xuân Trường
Preprocessing	Nguyễn Thanh Tuấn, Hoàng Xuân Trường
Model	Nguyễn Thanh Tuấn, Hoàng Xuân Trường
Test evalution, Visualize	Nguyễn Thanh Tuấn, Hoàng Xuân Trường
Slide	Nguyễn Thanh Tuấn, Hoàng Xuân Trường
README	Nguyễn Thanh Tuấn, Hoàng Xuân Trường

Contribute



Tổng kết

Kết quả: Nhóm đã học được 1 quy trình cơ bản của một data science. Các bước được thực hiện trong đồ án này:

- Crawl data
- Preprocessing
- Model and trainning
- Testing and visualize

Mặc dù kết qủa không được như mong đợi (độ chính xác không cao) nhưng đây là những bước cơ bản để có thể tự tìm hiểu thêm.