**Công trình liên quan**

Tên công trình:

Dự đoán giá nhà ở sử dụng phương pháp Linear Regression bằng Python( Predicting Housing Prices with Linear Regression using Python, pandas, and statsmodels )

Công trình này do tổ chức LearnDatasci nghiên cứu.

Tập dữ liệu: Housing Prices

Tập dữ liệu housing prices này có 74 mẫu (attributes) và 16 thuộc tính. Trong đó có 14 thuộc tính dùng để phân loại, 1 thuộc tính norminal là thuộc tính **day** và một thuộc tính cần dự đoán đó là housing price index.

Bảng định nghĩa các thuộc tính của tập dữ liệu:

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên biến | Định nghĩa | Tính chất |
| 1 | housing price index | Sự thay đổi giá nhà ở theo thời gian | Biến phụ thuộc |
| 2 | date | Cột chứa thông tin ngày tháng |  |
| 3 | season | Các mùa: fall, winter, spring, summer |  |
| 4 | gross domestic product | Thước đo tiền tệ thị trường | Khi giá trị này tăng thì giá nhà tăng |
| 5 | total expenditures | Tổng chi tiêu của Hoa Kì |
| 6 | producer price index | Thước đo đo lường giá trị sản phẩm |
| 7 | total unemployed | Tổng số người thất nghiệp |
| 8 | more than 15 weeks | Tổng số người thất nghiệp trên 15 tuần | Khi giá trị này tăng thì giá nhà giảm |
| 9 | not in labor searched for work | Số lường người không nằm trong độ tuổi lao động nhưng vẫn kiếm việc làm |
| 10 | multi jobs | Một người nhiều công việc |
| 11 | leavers | Những người muốn nghỉ việc và đang tìm việc làm khác |
| 12 | loser | Những người bị nghỉ việc |
| 13 | federal funds rate | lãi suất cho vay tại ngân hàng |
| 14 | sp500 (S&P 500) | Chỉ số chứng khoán Mỹ của 500 công ty niêm yết trên NYSE hoặc NASDAQ | Khi giá trị này tăng thì giá nhà tăng |
| 15 | consumer price index | Đo lường sự thay đổi mức độ hàng hóa tiêu dùng và dịch vụ |
| 16 | long interest rate | Lãi suất dài hạn của trái phiếu trong 10 năm | Khi giá trị này tăng thì giá nhà giảm |

Trong số các thuộc tính trên những giá trị được lựa chọn để phân tích:

Biến phụ thuộc (biến cần dự đoán): housing\_price\_index (HPI) là giá trị đo lường sự thay đổi giá nhà ở.

Các biến dùng để dự đoán được lựa chọn theo trực giác mà cho là ảnh hưởng lớn tới thị trường, hoạt động kinh tế như là: unemployment ( tỉ lệ thất nghiệp),  interest rates (lãi suất) , and gross domestic product (giá trị tiền tệ), total productivity (tổng lượng sản phẩm). Các thuộc tính này tác động khá lớn tới giá trị của nhà đất.

Phương pháp để giải quyết bài toán:

Do các thuộc tính của mẫu giá trị là các số thực (nummeric) nên ta dùng kỹ thuật OLS(Ordinary Least Squares Assumptions).

OLS là phương pháp bình phương nhỏ nhất dùng để đo độ chính xác của mô hình hồi quy tuyến tính. Trong bài toán là hồi quy tuyến tính đa biến.

Tỉ lệ thực nghiệm của công trình:

**Lấy dữ liệu từ web**

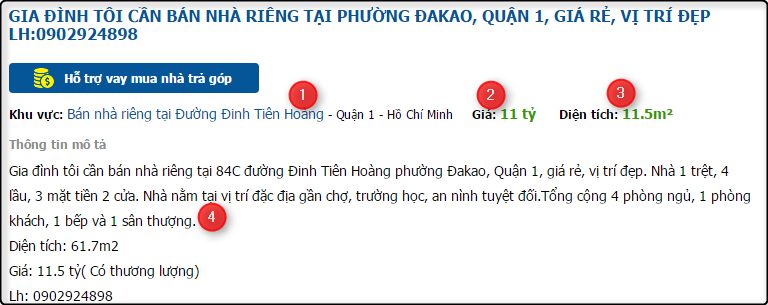
Khi lấy dữ liệu trên web: “batdongsan.com.vn” phần bán nhà riêng có hơn 17000 mẫu dữ liệu.

Phần dữ liệu dùng để phân tích thuộc loại:

|  |  |
| --- | --- |
| Loại nhà đất | Bán nhà riêng |
| Thành phố | Hồ Chí Minh |

Sử dụng thư viện Jsoup trong Java để phân tích tài liệu HTML: để lấy dữ liệu và thao tác dữ liệu

Link trang web của một mẫu ví dụ: <http://batdongsan.com.vn/ban-nha-rieng-duong-dinh-tien-hoang-phuong-da-kao/gia-toi-can-tai-dakao-quan-1-gia-re-vi-tri-dep-lh-0902924898-pr12356747>



Hình 1: Thông tin dữ liệu trên web rao bán nhà đất.



Hình 2: Thông tin đặc điểm bất động sản và cách liên hệ trên web rao bán nhà đất

Ghi chú: Các thuộc tính được rút trích từ thông tin trên web:

1. Địa chỉ
2. Giá tiền
3. Diện tích
4. Mô tả thông tin
5. Loại tin rao
6. Đường vào
7. Hướng nhà
8. Số tầng
9. Số phòng ngủ
10. Số toilet
11. Tên liên lạc
12. Địa chỉ liên lạc
13. Điện thoại
14. Mã tin đăng
15. Ngày đăng
16. Ngày hết hạn

Bảng các thuộc tính của dữ liệu:

|  |  |  |
| --- | --- | --- |
| STT | Tên thuộc tính | Mô tả thuộc tính |
| 1 | Địa chỉ | Địa chỉ của căn nhà |
| 2 | Giá tiền | Giá tiền của căn nhà (đơn vị triệu) |
| 3 | Diện tích | Diện tích của căn nhà (đơn vị m2) |
| 4 | Quận | Quận của căn nhà ở TP Hồ Chí Minh |
| 5 | Điện thoại | Là điện thoại của người rao bán căn nhà. |
| 6 | Mã số ĐK | Mã số đăng kí tin rao bán trên trang web. |
| 7 | Loại tin rao | Là loại tin rao, dữ liệu hiện thuộc loại bán nhà riêng. |
| 8 | Ngày đăng | Ngày đăng tin rao bán trên trang web. |
| 9 | Ngày hết hạn | Ngày tin rao bán hết hạn trên trang web. |
| 10 | Số tầng | Số tầng của căn nhà. |
| 11 | Số phòng | Số phòng của căn nhà. |
| 12 | Mặt tiền | Chiều rộng vỉa hè của căn nhà. |
| 13 | Đường vào | Độ dài đường dẫn vào căn nhà. |
| 14 | Số toilet | Số nhà vệ sinh của căn nhà. |
| 15 | Nội thất | Trang thiết bị nội thất trong căn nhà. |
| 16 | Thông tin người liên hệ | Thông tin người liên hệ như tên, điện thoại, địa chỉ của người rao bán |
| 17 | Mô tả | Mô tả thông tin của căn nhà. |
| 18 | Link bài đăng | Link bài đăng lên trang web. |

Bảng 1: Các thuộc tính của tập dữ liệu trước khi xử lý.

Ví dụ mẫu dữ liệu chưa xử lý

|  |  |  |
| --- | --- | --- |
| STT | Tên thuộc tính | Mô tả thuộc tính |
| 1 | Địa chỉ | **Đường Trần Hưng Đạo, Phường Nguyễn Cư Trinh, Quận 1, Hồ Chí Minh** |
| 2 | Giá tiền | **3.25 tỷ** |
| 3 | Diện tích | **100m²** |
| 4 | Quận | **Quận 1** |
| 5 | Điện thoại | **0911199995** |
| 6 | Mã số ĐK | **11772831** |
| 7 | Loại tin rao | **Bán nhà riêng** |
| 8 | Ngày đăng | **01-04-2017** |
| 9 | Ngày hết hạn | **08-04-2017** |
| 10 | Số tầng | **4 (tầng)** |
| 11 | Số phòng | **2(phòng)** |
| 12 | Mặt tiền | **3,00 (m)** |
| 13 | Đường vào |  |
| 14 | Số toilet | **4** |
| 15 | Nội thất |  |
| 16 | Thông tin người liên hệ | Tên Hùng.  SĐT: **0911199995** |
| 17 | Mô tả |  |
| 18 | Link bài đăng | **http://batdongsan.com.vn/ban-nha-rieng-duong-tran-hung-dao-phuong-nguyen-cu-trinh/chinh-chu-120-3-27-pham-ngu-lao-q1-moi-4-tang-3pn-4wc-chi-3-25-ty-pr11772831** |

Bảng 2: Một mẫu dữ liệu trước khi xử lý

Dữ liệu trước khi xử lý có hơn 17000 mẫu được lấy từ trên website mua bán nhà đất.

**Xử lý dữ liệu**

Các thuộc tính của tập dữ liệu:

|  |  |  |
| --- | --- | --- |
| STT | Tên thuộc tính | Mô tả thuộc tính |
|  | Giá tiền | Giá tiền của căn nhà (đơn vị triệu) |
|  | Diện tích | Diện tích của căn nhà (đơn vị m2) |
|  | Quận | Quận của căn nhà ở TP Hồ Chí Minh |
|  | Mã số ĐK | Mã số đăng kí tin rao bán trên trang web. |
|  | Số tầng | Số tầng của căn nhà. |
|  | Vị trí | Dữ liệu được rút trích từ phần mô tả. |

Bảng 3: Các thuộc tính của tập dữ liệu sau khi xử lý

Kiểu dữ liệu:

|  |  |  |
| --- | --- | --- |
| STT | Tên thuộc tính | Mô tả thuộc tính |
|  | Giá tiền | numeric |
|  | Diện tích | numeric |
|  | Quận | nominal |
|  | Mã số ĐK | nominal |
|  | Số tầng | numeric |
|  | Vị trí | nominal |

Bảng 4: Kiểu dữ liệu của tập cơ sở dữ liệu.

**Thuộc tính giá tiền:**

Các giá trị của thuộc tính này lấy tự web thuộc có các đơn vị như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Đơn vị | STT mẫu | Ví dụ | Sau khi xử lý | Đơn vị sau khi xử lý | Ghi chú |
| Triệu | 572 | 720 triệu | 720 | Triệu |  |
| Tỷ | 2 | 3.25 tỷ | 3250 | Triệu |  |
| Triệu/m2 | 276 | 3.8 triệu/m² | 123 | Triệu | DT:32.38m²Tổng giá tiền sau khi xử lý = 3.8 \* 32.38. |
| Thỏa thuận | 74 | Thỏa thuận | Loại bỏ dòng dữ liệu có giá trị thỏa thuận | |  |

Các mẫu có giá trị là thỏa thuận sẽ bị loại bỏ bởi vì thuộc tính giá tiền là thuộc tính cuối cùng, là thuộc tính dùng để dự đoán, nên nếu thuộc tính này mising thì không thể đưa vào tập dữ liệu để huấn luyện (train data) được.

Các giá trị còn lại có đơn vị khác nhau nên cần thống nhất lại thành một đơn vị cụ thể để có thể dễ dàng xử lý dữ liệu.

**Thuộc tính diện tích:**

Tương tự như thuộc tính giá tiền thì thuộc tính diện cũng có hai loại giá trị tồn tại đó là: mét vuông và không xác định.

Có một số mẫu do khi đưa lên thì giá trị của dữ liệu về phần diện tích có một số ảnh hưởng như người rao bán không biết chính xác phần diện tích như thế nào v.v, nên những mẫu này phần giá trị về diện tích sẽ mặc định là 0m2 hoặc có giá trị là “Không xác định”.

Một số mẫu có giá trị như trên là:

Mẫu 364, 2268 và một số mẫu khác có giá trị là 0m2

Mẫu 46, 64 và một số mẫu khác có giá trị là “Không xác định”.

Đối với những mẫu nào có hai giá trị như trên chúng ta cũng tiến hành loại bỏ để tập dữ liệu chính xác hơn.

**Thuộc tính quận:**

Thuộc tính này có kiểu dữ liệu là nominal gồm có 24 loại giá trị tương ứng 24 quận tại thành phố Hồ Chí Minh.

Bảng 24 quận tại TP Hồ Chí Minh lấy dữ liệu trên web và sau khi xử lý dữ liệu

|  |  |
| --- | --- |
| Dữ liệu ban đầu | Sau khi xử lý |
| Quận 1 | quan1 |
| Quận 2 | quan2 |
| Quận 3 | quan3 |
| Quận 4 | quan4 |
| Quận 5 | quan5 |
| Quận 6 | quan6 |
| Quận 7 | quan7 |
| Quận 8 | quan8 |
| Quận 9 | quan9 |
| Quận 10 | quan10 |
| Quận 11 | quan11 |
| Quận 12 | quan12 |
| Quận Bình Tân | binhtan |
| Quận Bình Thạnh | binhthanh |
| Quận Gò Vấp | govap |
| Huyện Hóc Môn | hocmon |
| Quận Tân Bình | tanbinh |
| Quận Tân Phú | tanphu |
| Quận Thủ Đức | thuduc |
| Huyện Bình Chánh | binhchanh |
| Huyện Nhà Bè | nhabe |
| Quận Phú Nhuận | phunhuan |
| Huyện Cần Giờ | cangio |
| Huyện Củ Chi | cuchi |

**Thuộc tính tầng**

Thuộc tính này có kiểu dữ liệu là numeric.

Cách xử lý dữ liệu của thuộc tính này:

|  |  |  |
| --- | --- | --- |
| **STT** | **Mẫu** | **Kết quả** |
| **1** | **4 (tầng)** | **4** |
| **2** | **7 (tầng)** | **7** |

Loại bỏ chữ tầng trong chuỗi dữ liệu để lấy giá trị là các số tự nhiên.

Các giá trị trong thuộc tính số tầng của tập dữ liệu chủ yếu nằm trong khoảng từ [1,9]. Những mẫu có giá trị khác khoảng này thì ta sẽ loại bỏ khỏi bộ dữ liệu:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Mẫu | Số lượng | Vị trí mẫu | Xử lý |
| 1 | 60 tầng | 1 | 9513 | Loại bỏ |
| 2 | 26 tầng | 1 | 3979 | Loại bỏ |
| 3 | 11 tầng | 2 | 206, 536 | Loại bỏ |
| 4 | 10 tầng | 3 | 589, 824, 1842 | Loại bỏ |

Những mẫu có giá trị không xác định (giá trị bỏ trống) cũng sẽ loại bỏ các mẫu này để bộ dữ liệu đưa vào phân tích được chính xác hơn.

Cụ thể có 1978 giá trị bi bỏ trống nên các mẫu này bị loại.

**Thuộc tính Mã ĐK**

Thuộc tính này có kiểu dữ liệu là nominal, mỗi mẫu có một giá trị khác với các mẫu còn lại, là giá trị duy nhất.

Nên thuộc tính này sẽ không được đưa vào để phân tích.

Thuộc tính Mã ĐK được tạo ra nhằm mục đích để truy vẫn thông tin từ tập dữ liệu gốc ban đầu.

**Thuộc tính vị trí**

Thuộc tính vị trí của kiểu dữ liệu là nominal.

Các giá trị của thuộc tính này sau khi được phân tích từ thuộc tính **Mô tả** của bảng dữ liệu ban đầu là:

Vị trí nhà hẻm.

Ví trị thuộc hẻm xe hơi.

Vị trí nhà mặt tiền.

Ví trí nhà có hai mặt tiền.

Các giá trị của thuộc tính này được sắp xếp theo cấp độ như bốn cấp độ trên.

Ta dùng phương thức biểu thức chính quy (Regular expressions) để tìm chuỗi thích hợp.

Để dùng biểu thức chính quy, từ **Mô tả**  tạo ra bộ từ điển để phân tích được.

Các từ trong bộ từ điển này được sắp xếp theo thứ tự dưới nhằm mục đích tìm kiểm chính xác hơn.

|  |  |  |
| --- | --- | --- |
| STT | Các từ | Giá trị sau khi phân tích |
| 1 | mặt tiền hxh | hemxehoi |
| 2 | mặt tiền hẻm xe hơi | hemxehoi |
| 3 | bán nhà hẻm xe hơi | hemxehoi |
| 4 | bán nhà hxh | hemxehoi |
| 5 | mt hxh | hemxehoi |
| 6 | bán gấp nhà hẻm xe hơi | hemxehoi |
| 7 | bán căn nhà hẻm xe hơi | hemxehoi |
| 8 | bán căn nhà hẻm xe tải | hemxehoi |
| 9 | bán gấp nhà hxh | hemxehoi |
| 10 | bán gấp nhà hẻm ô tô | hemxehoi |
| 11 | bán gấp nhà hẻm ôtô | hemxehoi |
| 12 | bán hẻm xe hơi | hemxehoi |
| 13 | bán nhà riêng hxh | hemxehoi |
| 14 | mặt tiền hẻm | hem |
| 15 | mt hẻm | hem |
| 16 | bán nhà hẻm | hem |
| 17 | bán nhà mặt tiền | hem |
| 18 | bán gấp nhà mt | mattien |
| 19 | bán nhà mt | mattien |
| 20 | bán nhà riêng măt tiền | mattien |
| 21 | bán căn nhà mặt tiền | mattien |
| 22 | bán gấp căn nhà mặt tiền | mattien |
| 23 | bán gấp căn nhà mt | mattien |
| 24 | bán nhà 2 mặt tiền | nhahaimattien |
| 25 | nhà hai mặt tiền | nhahaimattien |
| 26 | bán gấp nhà 2 mặt tiền | nhahaimattien |
| 27 | bán căn nhà 2 mặt tiền | nhahaimattien |
| 28 | nhà sau lưng mặt tiền | hem |
| 29 | bán nhà hẻm | hem |
| 30 | bán gấp nhà hẻm | hem |
| 31 | bán nhanh nhà hẻm | hem |
| 32 | nhà bán hẻm | hem |
| 33 | bán gấp căn nhà hẻm | hem |
| 34 | bán căn nhà hẻm | hem |
| 35 | căn nhà nằm trong hẻm | hem |
| 36 | bán nhà 2 mặt hẻm | hem |
| 37 | nhà hẻm | hem |
| 38 | hxh | hemxehoi |
| 39 | hẻm xe hơi | hemxehoi |
| 40 | nhà mặt tiền | mattien |
| 41 | nhà 2 mặt hẻm | hem |
| 42 | hẻm | hem |
| 43 | nhà cách mặt tiền | hem |

Bằng cách xử lý tự động bằng ngôn ngữ lập trình java.

Thuộc tính diện tích:

Dữ liệu gốc

Thuộc tính vị trí gồm những giá trị là: mặt tiền, hai mặt tiền, hẻm xe hơi, hẻm.

Những giá trị này được rút trích thông tin từ thuộc tính mô tả của bản dữ liệu gốc.

Để rút trích được những thông tin này, chúng ta tạo một bộ từ điển gồm những cụm từ có thể xuất hiện trong đoạn mô tả.

Bảng từ điển

Thuộc tính diện tích: có đơn vị là mét vuông. Kiểu dữ liệu của thuộc tính này là số thực.

Kỹ thuật OLS(Ordinary Least Squares Assumptions).

là phương pháp bình phương nhỏ nhất dùng để đo độ chính xác của mô hình hồi quy tuyến tính.

Phương pháp này được xây dựng dựa trên các giả định.

Hồi quy tuyến tính đơn(Simple Linear Regression) dùng để xác định biến phụ thuộc y dựa trên một biến độc lập x mà x liên quan một cách tuyến tính với biến y, theo phương trình như sau:

yi = α+ βxi + εi

Trong đó:

y là biến phụ thuộc

β hệ số tương quan

α hệ số chắn

x là giá trị độc lập được sử dụng để dự đoán y

ε hệ số sai số