# ỨNG DỤNG CÁC MÔ HÌNH LM, PCR, VÀ RF ĐỂ DỰ BÁO GIÁ NHÀ TRÊN ĐỊA BÀN THÀNH PHỐ HÀ NỘI

● PHẠM THỊ THANH HUYỀN

**TÓM TẮT:**

Nghiên cứu nhằm so sánh hiệu quả của 3 mô hình dự đoán giá nhà khác nhau: Mô hình hồi quy tuyến tính, PCR (Hồi quy thành phần chính), và Random Forest. Bằng cách sử dụng dữ liệu về bất động sản tại Hà Nội, nghiên cứu này đặt ra mục tiêu đánh giá khả năng dự đoán của từng mô hình dựa trên các biến độc lập như diện tích, số tầng, số phòng ngủ và các yếu tố khác. Kết luận của nghiên cứu nhấn mạnh vào sự quan trọng của lựa chọn mô hình phù hợp cho từng bộ dữ liệu cụ thể để đảm bảo tính chính xác và độ tin cậy trong dự đoán giá nhà hỗ trợ cho những người có nhu cầu tìm nhà trên địa bàn TP. Hà Nội trong giai đoạn này.

**Từ khóa:** mô hình, dự báo, giá cả, nhà đất, TP. Hà Nội, bất động sản.

**1. Đặt vấn đề**

Trong lĩnh vực bất động sản, việc định giá nhà đóng một vai trò quan trọng trong quá trình giao dịch và quyết định đầu tư. Đặc biệt, sau đại dịch Covid-19, giá nhà đất trên địa bàn Hà Nội liên tục biến động, gây khó khăn cho những người có nhu cầu thực sự về chỗ ở, cũng như các nhà đầu tư. Tuy nhiên, việc đưa ra một dự báo chính xác về giá trị của một căn nhà là một thách thức. Vì vậy, nghiên cứu về các phương pháp dự đoán giá nhà rất cần thiết để cung cấp thông tin hữu ích cho các bên liên quan trong quá trình mua bán và đầu tư bất động sản. Nghiên cứu này tập trung vào việc so sánh và đánh giá hiệu suất của các mô hình dự đoán giá nhà, bao gồm mô hình định giá Hedonic, mô hình hồi quy tuyến tính, mô hình PCR và mô hình Random Forest trên cơ sở tập dữ liệu về bất động sản tại Hà Nội.

**2. Phương pháp nghiên cứu**

***2.1. Nguồn dữ liệu***

Nghiên cứu này sử dụng dữ liệu từ Kaggle có tên Vietnam Housing Dataset (Hanoi) được đăng bởi tác giả Le Anh Duc. Dữ liệu gốc: 82.495 quan sát và 13 biến với cụ thể các biến gồm: STT, Ngày, Địa chỉ, Quận, Huyện, Loại hình nhà ở, Giấy tờ pháp lý, Số tầng, Số phòng ngủ, diện tích, dài, rộng và giá/m2. Trong nghiên cứu của mình, tác giả Le Anh Duc đã sử dụng mô hình ANN và RF để xác định các nhân tố ảnh hưởng đến giá nhà đất trên địa bàn Hà Nội. Tuy nhiên, những mô hình này có độ chính xác chưa cao. Bằng việc ứng dụng các mô hình LM, PCR và RF để dự báo giá nhà, kết quả dự báo sẽ có tính chính xác cao hơn.

***2.2. Tiền xử lí dữ liệu***

*2.2.1. Lọc dữ liệu*

Xử lý missing - value: Có 3 cách xử lý missing - value thường được sử dụng:

- Loại bỏ missing - value: Cách này dùng trong trường hợp tỉ lệ missing - value trong biến đó quá ít (khoảng dưới 3% so với tổng quan sát), hoặc missing - value không quan trọng đối với dữ liệu.

- Thay thế missing - value bằng một giá trị

mean/median/mode: Với cách này, ta điền những giá trị bị thiếu bằng giá trị mean hay median của một vài biến nếu biến liên tục và điền mode, nếu biến là biến categorical. Tuy nhiên, cách này có thể khiến phương sai của dữ liệu bị giảm xuống.

- Loại bỏ biến chứa missing - value: Cách này chỉ được sử dụng nếu tỉ lệ missing - value trong biến chiếm hơn 50% hoặc biến đó được đánh giá là không tác động đến biến mục đích hay những biến khác trong bộ dữ liệu.

Kế tiếp, chúng ta tiếp tục tiến hành loại bỏ các outliers, với 2 phương án khá phổ biến:

- Phương pháp IQR filtering: là một cách để lọc dữ liệu bằng cách sử dụng phạm vi của phần trung bình 50% của mẫu dữ liệu. Bằng cách này, chúng ta có thể loại bỏ các giá trị ngoại lai hoặc các giá trị cực đại/cực tiểu có thể ảnh hưởng đến tính toàn vẹn của phân tích dữ liệu. Điều này giúp cải thiện độ chính xác của phân tích và đánh giá chính xác hơn về phân phối của dữ liệu.

- Phương pháp Z-score filtering: là một kỹ thuật thống kê được sử dụng để chuẩn hóa và so sánh các điểm dữ liệu từ các phân phối khác nhau. Nó đo lường xem một điểm dữ liệu cách bao nhiêu độ lệch chuẩn so với giá trị trung bình của phân phối.

Qua quá trình chọn lọc, do số outliners là quá nhiều, có những outliners thậm chí đã để giá nhà lên tới cả trăm tỉ/m² và kể cả khi dùng IQR, vẫn còn các outliner lên tới cả tỉ/m², vậy nên đề án này sẽ sử dụng phương pháp [2] với hệ số z-score=0.7 để lọc 1 lượng lớn các outliners. Sau đây là thông số của dữ liệu sau khi loại bỏ các outliners. (Hình 1)

*2.2.2. Vẽ heatmap*

Dễ dàng quan sát thấy các biến như Dài, Rộng và Diện tích sẽ có tương quan nhất định tới nhau, hay các biến như Số tầng và Số phòng ngủ cũng là những nhóm các biến có thể tương quan tới nhau một cách chặt chẽ. Cùng xem bảng heatmap dưới đây để có thể hình dung sự tương quan nhau giữa các biến độc lập này là như thế nào.

Qua Heatmap, thấy được các biến đều có độ tương quan khá cao với nhau, điều này tương đối dễ hiểu và có thể giải thích được. Lí do: Số phòng ngủ và Số tầng có tương quan cao (cụ thể là 0.18) vì khi xây thêm tầng thì số phòng chức năng cũng sẽ tăng. Diện tích và số phòng ngủ cũng có độ tương quan lớn (0.38), khi có nhiều diện tích hơn, người ta cũng sẽ có xu hướng xây nhiều phòng ngủ hơn.

***2.3. Lựa chọn mô hình***

*2.3.1. Công cụ đánh giá mô hình*

+ Mean Squared Error (MSE) có lẽ là số liệu phổ biến nhất được sử dụng cho các bài toán hồi quy. Về cơ bản, nó tìm thấy sai số bình phương trung bình giữa các giá trị được dự đoán và thực tế. MSE là thước đo chất lượng của một công cụ ước tính - nó luôn không âm và các giá trị càng gần 0 càng tốt.*

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

***Hình 1: Bộ dữ liệu trước và sau khi lọc***

*(a) Bộ dữ liệu trước khi lọc*

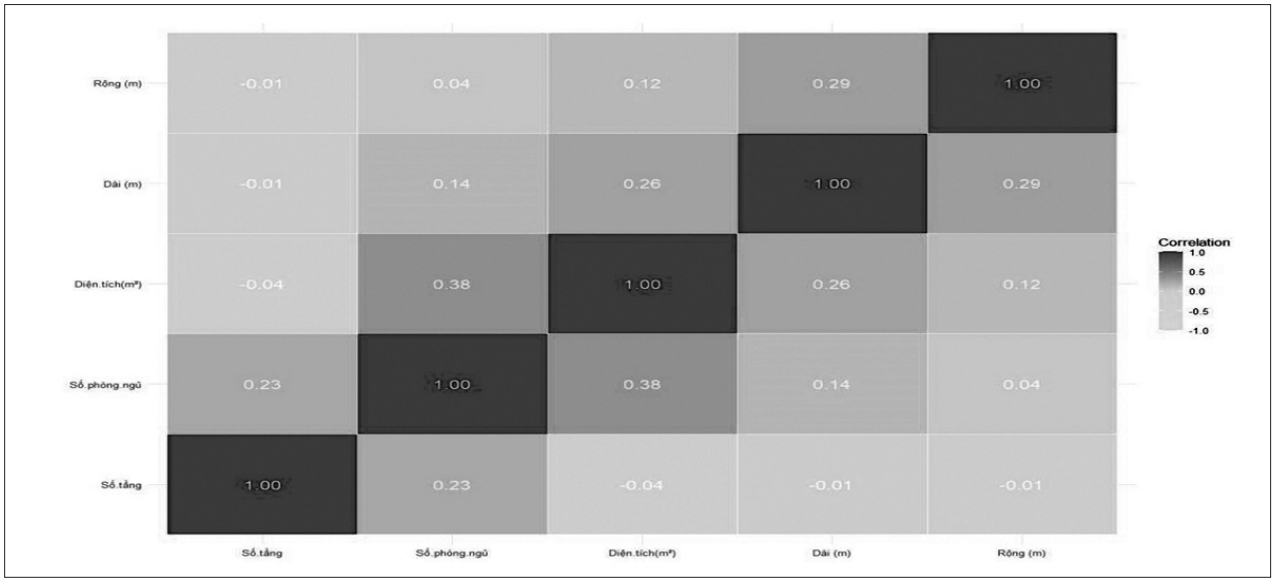
| | X | Ngày | Địa.chỉ | Quận | Huyện | Loại.hình.nhà.ở |
|---|---|---|---|---|---|---|
| | Min. : 2 | Length:11473 | Length:11473 | Length:11473 | Length:11473 | Length:11473 |
| | 1st Qu.:21279 | Class :character | Class :character | Class :character | Class :character | Class :character |
| | Median :41611 | Mode :character | Mode :character | Mode :character | Mode :character | Mode :character |
| | Mean :41684 | | | | | |
| | 3rd Qu.:62260 | | | | | |
| | Max. :82362 | | | | | |
| | Giấy.tờ.pháp.lý | Số.tầng | Số.phòng.ngủ | Diện.tích(m²) | Dài (m) | Rộng (m) |
| | Length:11473 | Min. : 1.000 | Min. : 1.000 | Min. : 1.00 | Min. : 1.00 | Min. : 1.00 |
| | Class :character | 1st Qu.: 4.000 | 1st Qu.: 3.000 | 1st Qu.: 34.00 | 1st Qu.: 8.00 | 1st Qu.: 3.80 |
| | Mode :character | Median : 4.000 | Median : 4.000 | Median : 40.00 | Median : 10.00 | Median : 4.00 |
| | | Mean : 4.432 | Mean : 3.955 | Mean : 55.05 | Mean : 45.42 | Mean : 17.64 |
| | | 3rd Qu.: 5.000 | 3rd Qu.: 4.000 | 3rd Qu.: 52.00 | 3rd Qu.: 12.00 | 3rd Qu.: 5.00 |
| | | Max. :73.000 | Max. :10.000 | Max. :40000.00 | Max. :120000.00 | Max. :5500.00 |
| | Giá.m2(triệu) | | | | | |
| | Min. : 1.0 | | | | | |
| | 1st Qu.: 71.2 | | | | | |
| | Median : 90.0 | | | | | |
| | Mean : 2000.2 | | | | | |
| | 3rd Qu.: 116.7 | | | | | |
| | Max. :934959.0 | | | | | |

*(b) Bộ dữ liệu sau khi lọc*

| | X | Ngày | Địa.chỉ | Quận | Huyện | Loại.hình.nhà.ở |
|---|---|---|---|---|---|---|
| | Min. : 2 | Length:8428 | Length:8428 | Length:8428 | Length:8428 | Length:8428 |
| | 1st Qu.:20972 | Class :character | Class :character | Class :character | Class :character | Class :character |
| | Median :41022 | Mode :character | Mode :character | Mode :character | Mode :character | Mode :character |
| | Mean :41279 | | | | | |
| | 3rd Qu.:61636 | | | | | |
| | Max. :82362 | | | | | |
| | Giấy.tờ.pháp.lý | Số.tầng | Số.phòng.ngủ | Diện.tích(m²) | Dài (m) | Rộng (m) | Giá.m2(triệu) |
| | Length:8428 | Min. :4.000 | Min. : 1.000 | Min. : 1.00 | Min. : 1.00 | Min. : 1.000 | Min. : 1.00 |
| | Class :character | 1st Qu.:4.000 | 1st Qu.: 3.000 | 1st Qu.: 33.00 | 1st Qu.: 8.00 | 1st Qu.: 3.600 | 1st Qu.: 73.68 |
| | Mode :character | Median :5.000 | Median : 4.000 | Median : 40.00 | Median : 10.00 | Median : 4.000 | Median : 90.28 |
| | | Mean :4.549 | Mean : 3.972 | Mean : 44.68 | Mean : 10.93 | Mean : 5.109 | Mean :101.23 |
| | | 3rd Qu.:5.000 | 3rd Qu.: 4.000 | 3rd Qu.: 50.00 | 3rd Qu.: 12.00 | 3rd Qu.: 5.000 | 3rd Qu.:111.76 |
| | | Max. :5.000 | Max. :10.000 | Max. :350.00 | Max. :185.00 | Max. :80.000 | Max. :950.00 |

*Nguồn: Tác giả thực hiện*

*Hình 2: Ma trận tương quan*




*Nguồn: Tác giả thực hiện*

Trong đó:

• $n$ là số lượng quan sát trong tập dữ liệu kiểm tra.

• $y_i$ là giá trị thực tế của quan sát thứ i.

• $\hat{y}_i$ là giá trị dự đoán của mô hình cho quan sát thứ $i$.

+ RMSE (Root Mean Squared Error) là căn bậc 2 của MSE, có công thức:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Trong đó:

• RMSE là Root Mean Squared Error.

• $n$ là số lượng quan sát trong dữ liệu.

• $y_i$ là giá trị thực tế của quan sát thứ i.

• $\hat{y}_i$ là giá trị dự đoán của mô hình cho quan sát thứ $i$.

Chỉ số MSE hay RMSE có giá trị thấp cho thấy mô hình có thể dự đoán giá trị mục tiêu với độ chính xác cao, trong khi chỉ số MSE hay RMSE cao sẽ cho thấy sự chênh lệch lớn giữa các giá trị dự đoán và thực tế.

*2.3.2. Mô hình định giá Hedonic*

Mô hình định giá hedonic (hedonic pricing model) bắt nguồn từ lý thuyết về giá trị do Lancaster (1966), Griliches (1971) và Rosen (1974) phát triển thông qua những nghiên cứu vào những năm 1960 - 1970. Lý thuyết này cho rằng một hàng hóa có vô số các thuộc tính kết hợp lại thành các nhóm thuộc tính và cấu thành nên giá của hàng hóa đó. Lancaster đã xây dựng nền tảng lý thuyết cho mô hình hedonic thông qua lý thuyết tiêu dùng năm 1966 rằng sự thỏa dụng của người tiêu dùng có được từ những đặc tính của sản phẩm, chứ không phải trực tiếp từ sản phẩm đó. Trong dự báo bất động sản, mô hình Hedonic sẽ giả định rằng giá của một căn nhà, sẽ là tổng các giá trị các đặc tính của nó như diện tích, vị trí, tiện ích xung quanh, chất lượng xây dựng, và các yếu tố khác. Để áp dụng mô hình Hedonic, thông thường người ta sẽ sử dụng các phương pháp hồi quy để ước tính giá trị của từng thuộc tính, sau đó kết hợp chúng lại để tính toán được giá trị dự đoán của sản phẩm. Ngoài thị trường bất động sản, mô hình Hedonic có thể được sử dụng trong nhiều lĩnh vực khác như thị trường hàng hóa tiêu dùng, giúp dự báo giá cả và hiểu rõ hơn về cách mà các đặc tính của một sản phẩm ảnh hưởng đến giá trị của nó.

*2.3.3. Mô hình hồi quy tuyến tính*

Về bản chất, mô hình hồi quy tuyến tính hay Linear Regression là một phương pháp thống kê cơ bản được sử dụng để tìm hiểu mối quan hệ tuyến tính giữa một biến phụ thuộc y (outcome) và một hoặc nhiều biến độc lập x (predictors). Mục tiêu

của Linear Regression là tạo ra một mô hình tuyến tính dựa trên dữ liệu để dự đoán giá trị của biến phụ thuộc khi biết giá trị của các biến độc lập.

Theo đó, mô hình giả định rằng mối quan hệ giữa các biến là tuyến tính, có nghĩa là biến phụ thuộc biến đổi tuyến tính theo các biến độc lập. Mô hình Linear Regression được biểu diễn bằng phương trình tuyến tính như sau:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$

Trong đó:

$y$: Biến phụ thuộc (outcome) mà chúng ta muốn dự đoán.

$x_1, x_2,...,x_n$: Các biến độc lập (predictors).

$\beta_0, \beta_1, \beta_2, ... \beta_n$: Các hệ số (coefficients) của mô hình, biểu thị độ ảnh hưởng của các biến độc lập đến biến phụ thuộc.

$\varepsilon$ : Sai số ngẫu nhiên, biểu thị sự biến thiên không thể giải thích được của biến phụ thuộc.

Trong quá trình xây dựng mô hình Linear Regression, chúng ta sử dụng các phương pháp như phương pháp OLS (Ordinary Least Squares - OLS) để ước lượng các hệ số. Mục tiêu của OLS là tìm ra mô hình mà làm giảm thiểu tổng bình phương của các sai số giữa giá trị dự đoán và giá trị quan sát thực tế.

Linear Regression được sử dụng rộng rãi trong nhiều lĩnh vực, từ nghiên cứu khoa học đến dự đoán tài chính, vì tính đơn giản và dễ hiểu của nó cũng như khả năng ứng dụng trong nhiều bài toán dự đoán và phân tích dữ liệu. Trong mô hình dự báo giá nhà, chúng ta sẽ dùng y hay biến phụ thuộc tương ứng với biến Giá/$m^2$ và tất các biến dạng số (numeric) còn lại như Diện tích, Số tầng, Số phòng ngủ,... là x hay biến độc lập để giải thích cho y.

*2.3.4. Mô hình PCR*

PCA (Principal Components Analysis): Phân tích thành phần chính là một phương pháp biến đổi giúp giảm số lượng lớn các biến có tương quan với nhau thành tập ít các biến sao cho các biến mới tạo ra là tổ hợp tuyến tính của những biến cũ không có tương quan lẫn nhau. Đặc tính của PCA như sau:

*1. Giảm số chiều dữ liệu:* PCA giúp giảm số chiều của dữ liệu, đặc biệt hữu ích khi dữ liệu có quá nhiều chiều thông tin, giúp cho việc trực quan hóa dữ liệu trở nên dễ dàng hơn.

*2. Tối ưu hóa sự biến thiên:* Bằng cách xoay trục tọa độ, PCA tạo ra một hệ trục tọa độ mới sao cho độ biến thiên của dữ liệu được tối đa hóa, giữ lại nhiều thông tin nhất mà không ảnh hưởng tới chất lượng của các mô hình dự báo.

*3. Xây dựng biến factor mới:* Trong không gian mới, PCA tạo ra những biến factor mới là tổ hợp tuyến tính của các biến ban đầu, giúp tăng khả năng hiểu và diễn giải dữ liệu.

*4. Khám phá thông tin mới:* Trong không gian mới, PCA có thể giúp chúng ta khám phá thêm những thông tin quý giá mới mà trong chiều thông tin cũ, những thông tin này có thể bị che mất.

Sử dụng prcomp để quan sát, ta có các bảng thành phần chính như sau:

Nhận xét: với 3 thành phần PC1, PC2. PC3 giải thích được xấp xỉ 76% độ biến động của mô hình, vậy mô hình hồi quy PCR của ta sẽ sử dụng 3 thành phần này làm thành phần chính.

*2.3.5. Mô hình Random Forest*

Random forest là thuật toán có thể giải quyết cả bài toán regression và classification. Thuật toán Random Forest hoạt động bằng cách xây dựng nhiều cây quyết định ngẫu nhiên trong quá trình huấn luyện. Mỗi cây trong rừng (forest) được xây dựng dựa trên một mẫu dữ liệu con ngẫu nhiên từ

***Hình 3: Bảng thành phần chính theo Prcomp***

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5
Standard deviation     1.2719 1.0866 0.9755 0.8419 0.7357
Proportion of Variance 0.3236 0.2361 0.1903 0.1418 0.1082
Cumulative Proportion  0.3236 0.5597 0.7500 0.8918 1.0000
```

*Nguồn: Tác giả thực hiện*

*Hình 4: Mô hình Random Forest*

```
'data.frame':   4214 obs. of  9 variables:
 $ Quận           : Factor w/ 23 levels "Huyện Chương Mỹ",..: 9 15 18 15 18 19 22 17 22 13 ...
 $ Loại.hình.nhà.ở: Factor w/ 4 levels "Nhà biệt thự",..: 3 3 3 3 3 3 3 2 3 2 ...
 $ Giấy.tờ.pháp.lý: Factor w/ 2 levels "có","không có": 1 1 1 1 1 1 1 1 1 1 ...
 $ Số.tầng        : num  4 4 5 4 5 5 4 4 5 5 ...
 $ Số.phòng.ngủ   : num  3 4 4 4 4 3 4 3 3 4 ...
 $ Diện.tích(m²)  : num  24 38 55 36 40 60 62 20 37 58 ...
 $ Dài (m)        : num  8 9 11 10 8 12 15 6 8 14 ...
 $ Rộng (m)       : num  3.4 4 5 3.6 5 5 4 4 4 4 ...
 $ Giá.m2(triệu)  : num  75 60.5 109.1 75 67.5 ...
```

*Nguồn: Tác giả thực hiện*

tập dữ liệu huấn luyện và sử dụng một số biến đặc trưng ngẫu nhiên để tạo ra sự đa dạng giữa các cây. Khi có dữ liệu mới, mỗi cây trong rừng sẽ đưa ra một dự đoán, sau đó dự đoán cuối cùng sẽ được tính toán bằng cách lấy trung bình của các dự đoán từ tất cả các cây. (Hình 4)

**3. Kết quả**

***3.1. Mô hình hồi quy thông thường***

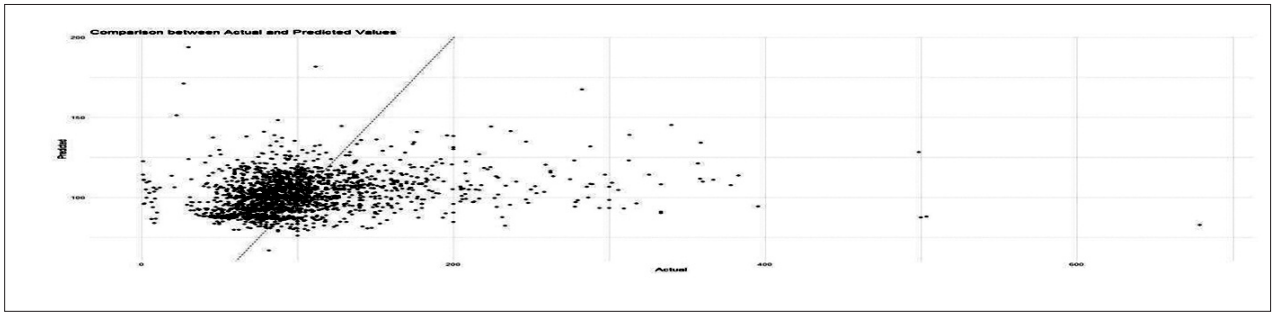
Hình 5 mô tả đồ thị dải điểm với trục Y là giá trị dự báo, trục X là giá trị thật của mô hình Hồi quy, dễ dàng quan sát được rằng các điểm phân phối lệch khá cao so với đường kẻ (độ dốc = 1). Ngoài ra, dựa vào việc các điểm có xu hướng lệch về bên phía trái đường kẻ nhiều hơn, có thể kết luận mô hình thường cho ra kết quả dự đoán cao hơn so với thực tế.

Nhận xét: Dựa vào $\beta$(coefficients), dường như diện tích và độ dài, rộng của căn nhà không ảnh hưởng lớn đến giá trị của căn nhà lắm, tuy nhiên số tầng và phòng ngủ thì có. Điều này có thể giải thích phần nào bởi tâm lý mua nhà tại Việt Nam với những miếng đất diện tích lớn sẽ có ít nhà đầu tư có nhu cầu hơn và đa phần mọi người đều sẽ chỉ có đủ điều kiện tài chính cho 1 căn nhà với diện tích không quá rộng nhưng lại đầy đủ tiện nghi (nhiều tầng và phòng)

Qua mô hình hồi quy tuyến tính thông thường, có thể rút được ra một vài kết luận như sau, hệ số $\beta$(coefficients) của "Dài" là âm, nhưng 2 biến có tương quan khá cao với "Dài" là "Diện tích" và "Rộng" lại có hệ số $\beta$(coefficients) dương, điều này dường như là vô lý. Ngoài ra, Hệ số R hiệu chỉnh chỉ xấp xỉ 0.04 hay 4%, điều này có nghĩa các biến độc lập ít giải thích được cho biến phụ thuộc (Giá/$m^2$). Mặt khác, p-value < 2.2e-16 (p-value thấp) cho thấy mô hình có một hoặc một số biến độc lập không có ảnh hưởng lớn đến biến phụ thuộc, nhưng khi kết hợp lại cùng nhau, chúng vẫn tạo ra một ảnh hưởng có ý nghĩa thống kê tốt. Kết luận lại, trong trường hợp này, dù mô hình có ý nghĩa thống kê, nhưng khả năng dự đoán của nó có thể không cao do $R^2$ thấp.

Những dòng cuối cùng mô hình này biểu thị khoảng cách kết quả dự đoán và chỉ số RMSE, theo đó với chỉ số RMSE lên tới 49,51 (50% so với trung bình giá nhà dự đoán, cụ thể là Mean = 101.83), ta kết luận rằng mô hình này không mang lại độ chính xác cao cho bộ dữ liệu. (Hình 6)

*Hình 5: Đồ thị dải điểm theo mô hình hồi quy*




*Nguồn: Tác giả thực hiện*

*Hình 6: Mô hình hồi quy*

```
Call:
lm(formula = `Giá.m2(triệu)` ~ ., data = train_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-170.07   -26.22   -10.06    10.30   866.60

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      17.98493    6.51924   2.759  0.00582 **
Số.tầng          12.50981    1.42088   8.804  < 2e-16 ***
Số.phòng.ngủ      3.75884    0.66543   5.649 1.69e-08 ***
`Diện.tích(m²)`   0.27721    0.03589   7.724 1.30e-14 ***
`Dài (m)`        -0.13736    0.08536  -1.609  0.10763
`Rộng (m)`        0.15011    0.12997   1.155  0.24816
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.27 on 6315 degrees of freedom
Multiple R-squared:  0.03963,   Adjusted R-squared:  0.03887
F-statistic: 52.12 on 5 and 6315 DF,  p-value: < 2.2e-16

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  66.85   93.37  100.75  101.83  108.38  193.78
[1] 49.51074
```

*Nguồn: Tác giả thực hiện*

***3.2. Mô hình PCR***

Hình 7 mô tả đồ thị dải điểm với trục Y là giá trị dự báo, trục X là giá trị thật của mô hình PCR, không quá khác biệt so với Hình 5 mô tả cho mô hình Hồi quy thông thường. Tuy nhiên, thấy lần này các điểm thường tụ lại theo từng cụm hơn thay vì rải rác. Các giá trị thực tế vẫn thường nhỏ hơn giá trị dự báo.

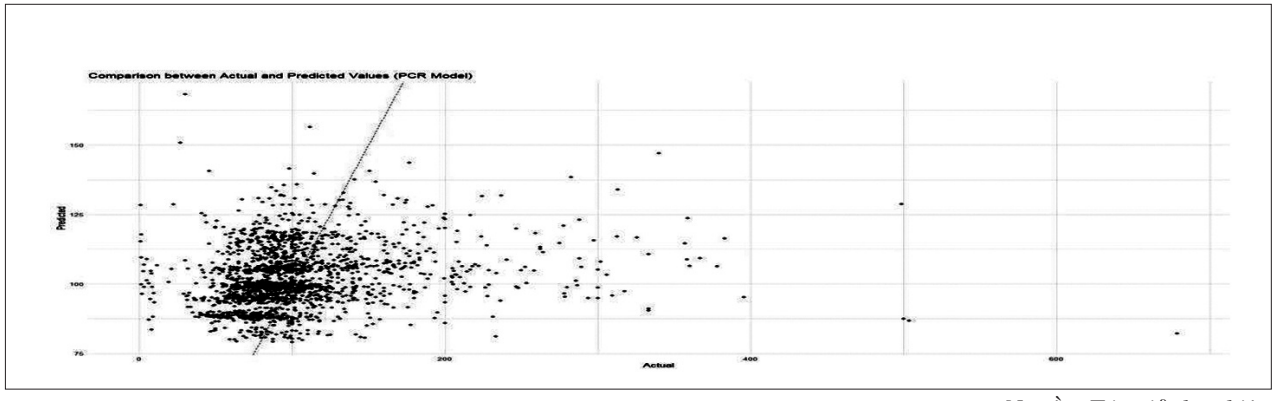
Như đã nói ở trên, PCR sẽ được xây dựng với 3 components PC1, PC2, và PC3, mô hình dự báo cho ra RMSE là 49.66, thậm chí còn cao hơn so với mô hình hồi quy tuyến tính ban đầu, dù không đáng kể. Tính không hiệu quả của mô hình PCR có thể được giải thích với lí do là có quá ít chiều của biến X ngay từ đầu, và độ tương quan giữa các biến độc lập chưa đủ cao. Kết luận lại mô hình PCR không phù hợp cho dữ liệu này. (Hình 8)

***3.3. Mô hình Random Forest***

Đối với mô hình Random Forest này, ta không cần chỉ dùng 5 biến có dạng numeric như 2 mô hình trên, thay vào đó chúng ta phân loại chúng theo 4 biến dạng categorical nữa nhờ sự ưu việt của mô hình này.

Hình 9 mô tả đồ thị dải điểm với trục Y là giá trị

*Hình 7: Đồ thị dải điểm theo mô hình PCR*




*Nguồn: Tác giả thực hiện*

*Hình 8: Mô hình dự báo PCR*

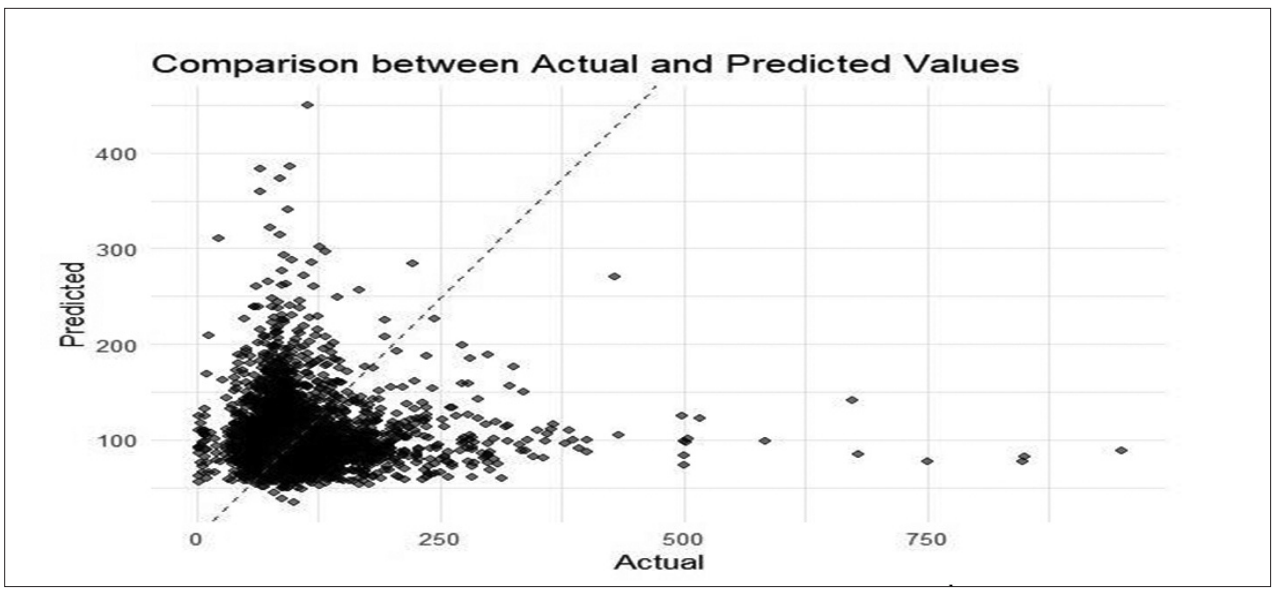
```
Data:   X dimension: 6321 5
        Y dimension: 6321 1
Fit method: svdpc
Number of components considered: 5

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
CV           55.36    54.79    54.36    54.37    54.35    54.31
adjCV        55.36    54.79    54.36    54.37    54.35    54.31

TRAINING: % variance explained
                1 comps  2 comps  3 comps  4 comps  5 comps
X                33.153   57.082   76.058   89.626  100.000
Giá.m2(triệu)     2.083    3.638    3.665    3.761    3.963
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  79.12   94.93   99.68  101.78  107.23  168.26
[1] 49.66287
```

*Nguồn: Tác giả thực hiện*

*Hình 9: Đồ thị dải điểm theo mô hình Random Forest*




*Nguồn: Tác giả thực hiện*

dự báo, trục X là giá trị thật của mô hình Random Forest, được phân bổ khác hoàn toàn so với đồ thị dải điểm của 2 mô hình trước đó. Lần này mô hình Random Forest không còn dự báo các giá trị cao hơn hẳn so với giá trị Actual nữa, song tuy nhiên chúng lại dự báo ra rất nhiều các giá trị Predict với sự chênh lệch lớn hơn hẳn so với giá trị Actual. Ngoài ra, do Random Forest cũng là thuật toán dựa báo trên Classification nên ta thấy các điểm được dự báo dường như được phân thành 2 loại đối xứng nhau qua đường phân chia.

Mô hình Random Forest này sẽ có số lượng cây là 500 và tương ứng với nó, số lượng biến được thử nghiệm tại mỗi phân chia là 2.

1. Mean squared residuals (Sai số bình phương trung bình): Đây là giá trị trung bình của bình

*Hình 10: Mô hình Random Forest*

```
Call:
 randomForest(formula = train_set$`Giá.m2(triệu)` ~ train_set$Loại.hình.nhà.ở +      train_set$Số.tầng + train_set$Quận +
train_set$Số.phòng.ngủ +      train_set$`Diện.tích(m²)` + train_set$`Dài (m)` + train_set$`Rộng (m)`,      data = train_set,
importance = T)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 2031.816
                    % Var explained: 25.94
                                    %IncMSE IncNodePurity
train_set$Loại.hình.nhà.ở 44.497411     1323430.1
train_set$Số.tầng         16.620322      358274.6
train_set$Quận            44.071390     1811213.0
train_set$Số.phòng.ngủ     7.871292      744749.4
train_set$`Diện.tích(m²)` 15.597363     2010606.1
train_set$`Dài (m)`       17.207431     1131638.9
train_set$`Rộng (m)`      10.607594     1016795.9
[1] 64.74504
```

*Nguồn: Tác giả thực hiện*

phương của sai số, tức là độ lệch giữa giá trị dự đoán và giá trị thực tế, trên tập huấn luyện. Trong trường hợp này, giá trị là 2031.816, cho thấy mức độ sai số trung bình trên tập huấn luyện.

2. % Var explained (Phần trăm biến thiên được giải thích): Đây là phần trăm của sự biến thiên trong biến phụ thuộc mà mô hình giải thích được. Trong trường hợp này, mô hình giải thích được khoảng 25.94% của sự biến thiên trong biến phụ thuộc.

Cuối cùng chỉ số RMSE là 64.74, không khó để thấy rằng sai số của mô hình Random Forest sẽ cao hơn so với 2 mô hình trước khi mà ở Hình 9, số các giá trị dự báo "Predict" có giá trị lớn hơn nhiều giá trị thật "Actual" là khá nhiều. (Hình 10)

**4. Kết luận**

Mục đích của nghiên cứu để tìm hiểu và so sánh hiệu quả của các mô hình học máy khác nhau trong việc dự đoán giá nhà sử dụng dữ liệu về bất động sản tại Hà Nội, các mô hình mô hình được sử dụng gồm Linaear Regression, PCR, và Random Forest Regression. Sai số của các mô hình theo tuần tự LM, PCR, FR lần lượt là: 49,51; 49,66 và 64,74, qua kết luận có thể thấy mô hình mang lại dự báo cách chính xác nhất cho dữ liệu là mô hình Linear Model và mô hình kém hiệu quả nhất là Random Forest với chỉ số RMSE cao hơn xấp xỉ 30% so với 2 mô hình Linear Model và PCR. Tuy nhiên, trong tương lai, với các bộ dữ liệu có thể tích hợp dữ liệu bất động sản từ các khu vực địa lý khác nhau cùng nhiều biến hơn (Số phòng khách, tuổi nhà) để tạo ra một mô hình đa dạng hơn. Mô hình PCR và Random Forest sẽ mang lại nhiều giá trị hơn so với tập dữ liệu hiện tại ■

**TÀI LIỆU THAM KHẢO:**

1. Hong Ru Chan (2024). Rent Price Prediction with Advanced Ma-chine Learning Methods: A Comparison of California and Texas. Highlights in Science, Engineering and Technology, 85.

2. Jengei Hong, Heey-oul Choi, Woo-sung Kim (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. International Journal of Strategic Property Management, 24(3).

3. Le Anh Duc (2022). Predicting Hanoi, Vietnam housing price using Artificial Neural Network (ANN) and Random Forest (RF) models for Regression. Predicting Hanoi housing price (ANN & RF) (kaggle.com).

4. Rosen S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 82, 34-55.

5. Lancaster K. J. (1966). A New Approach to Consumer Theory. Journal of Political Economy, 74, 132-157.
6. Griliches Z. (1971). Price Indexes and Quality Change. Harvard University Press, Cambridge.

**Ngày nhận bài: 21/3/2024**
**Ngày phản biện đánh giá và sửa chữa: 5/4/2024**
**Ngày chấp nhận đăng bài: 23/4/2024**

*Thông tin tác giả:*
**ThS. PHẠM THỊ THANH HUYỀN**
**Khoa Kế toán - Tài chính, Trường Đại học Hải Phòng**

# USING LINEAR REGRESSION MODEL, PRINCIPAL COMPONENTS REGRESSION (PCR), AND RANDOM FOREST TO PREDICT HOUSE PRICES IN HANOI

● Master. **PHAM THI THANH HUYEN**
Faculty of Accounting - Finance, Hai Phong University

**ABSTRACT:**
This study compared the effectiveness of three different house price prediction models, including the linear regression model, principal components regression (PCR), and random forest. Using real estate data in Hanoi, this study aimed to evaluate the predictability of each model based on independent variables, such as square footage, number of floors, number of bedrooms, and other factors. This study's conclusion emphasized the importance of choosing the appropriate model for each specific data set to ensure accuracy and reliability in predicting house prices in Hanoi.

**Keywords:** model, forecast, prices, real estate, Hanoi.