Đồ án cuối kỳ: Giá căn hộ cho thuê tại Mỹ

Lê Quang Trung, Trần Hoài Bắc,
Tạ Hoàng Kim Thy, Nguyễn Hải Ngọc Huyền,
Lê Hồ Hoàng Anh, Nguyễn Phúc Loan

Ngày 28 tháng 7 năm 2024

Outline

Giới thiệu chung

- Về dataset được sử dụng: Đây là một tập dữ liệu về danh sách các căn hộ cho thuê tại Mỹ từ nhiều nền tảng môi giới bất động sản khác nhau. Tập dữ liệu bao gồm 10.000 bản ghi cho thuê với 22 cột. Dữ liệu có chứa các giá trị thiếu nhưng đã được làm sạch theo cách mà các cột price và square_feet không bao giờ bị bỏ trống, trong khi đó, tập dữ liệu vẫn được lưu giữ nguyên vẹn như lúc ban đầu.
- Trong bối cảnh thành phố đông đúc với nhu cầu thuê căn hộ tăng cao, việc định giá căn hộ trở nên thách thức.
- Bài toán đặt ra của phân tích này là xây dựng được một mô hình dự đoán giá thuê căn hộ chính xác, hỗ trợ người cho thuê và môi giới định giá căn hộ hợp lý, phù hợp với nhu cầu thị trường.

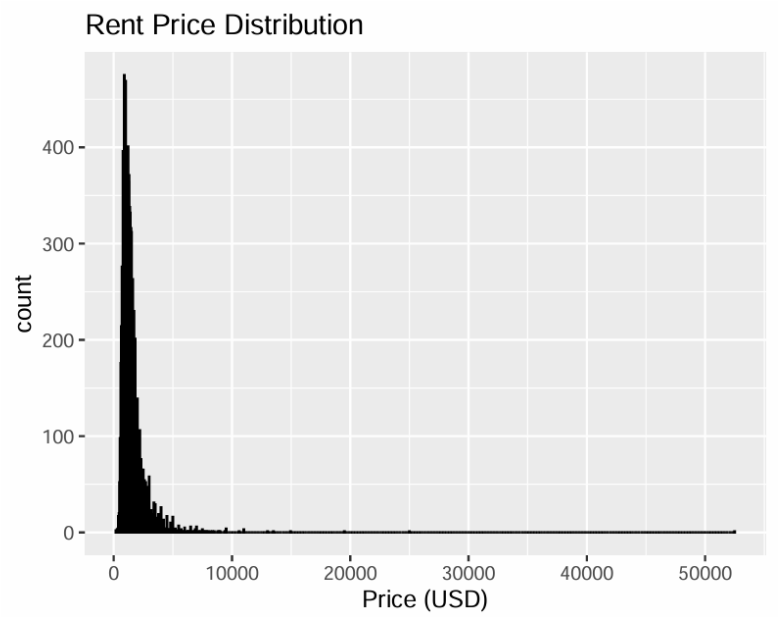
Đề xuất các phương pháp xử lý dữ liệu

- Phân tích thống kê và trực quan hóa dữ liệu
- Tiền xử lý dữ liệu
  - Xử lý giá trị thiếu
  - Chuyển đổi định dạng dữ liệu
  - Chuẩn hóa dữ liệu
  - Xử lý giá trị ngoại lai
- Xây dựng và đánh giá mô hình
- Kiểm tra và tối ưu hóa mô hình
- Kết quả và ứng dụng

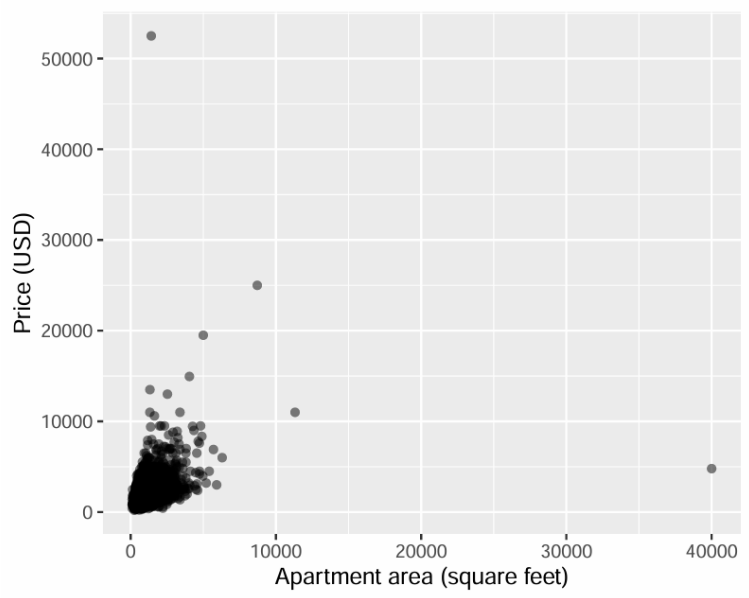
Các mục tiêu cần đạt được

- Xử lý dữ liệu hiệu quả
- Biến đổi dữ liệu để cải thiện phân phối
- Xử lý các biến phân loại
- Phân tích và loại bỏ giá trị ngoại lai
- Phân tích tương quan và quan hệ giữa các biến
- Xây dựng và đánh giá mô hình dự đoán
- Tối ưu hóa mô hình và đưa ra dự đoán chính xác
- Cung cấp những hiểu biết quan trọng về thị trường

EDA

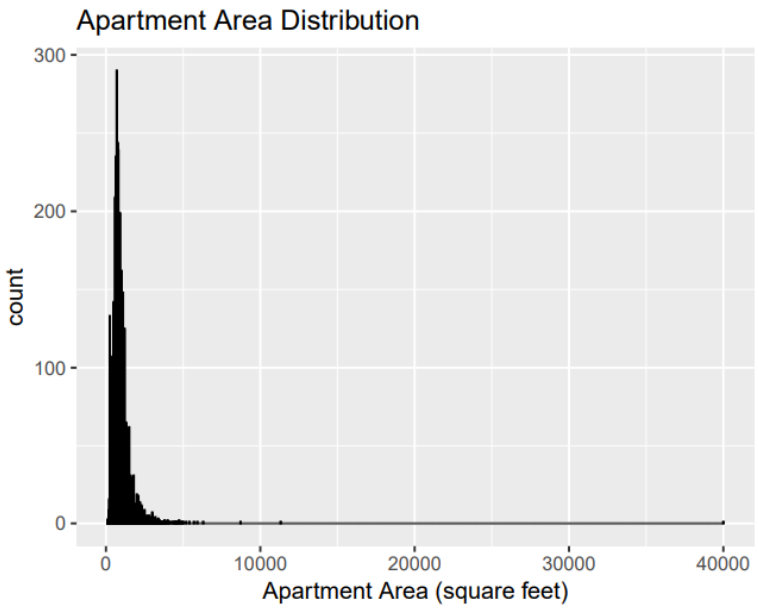


Hình: Phân phối giá thuê

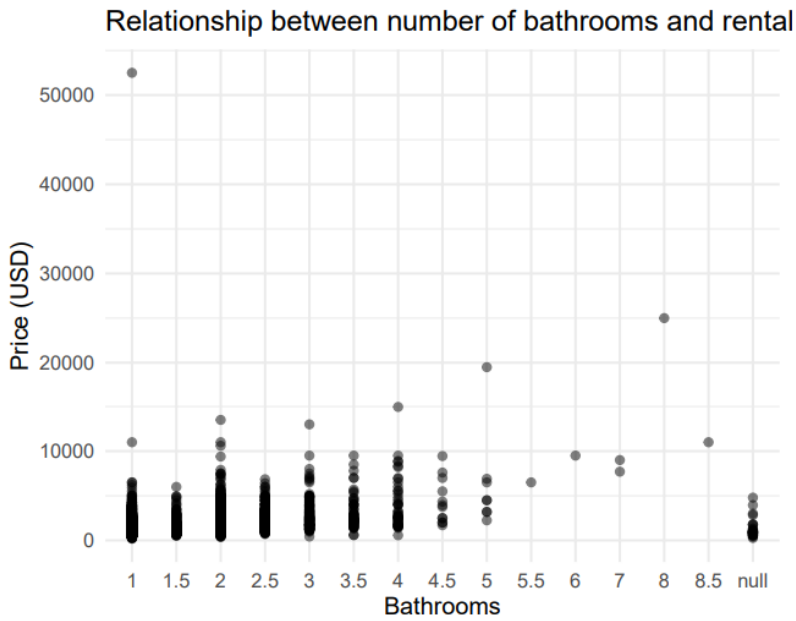


Hình: Mối quan hệ giữa diện tích căn hộ và giá thuê

EDA

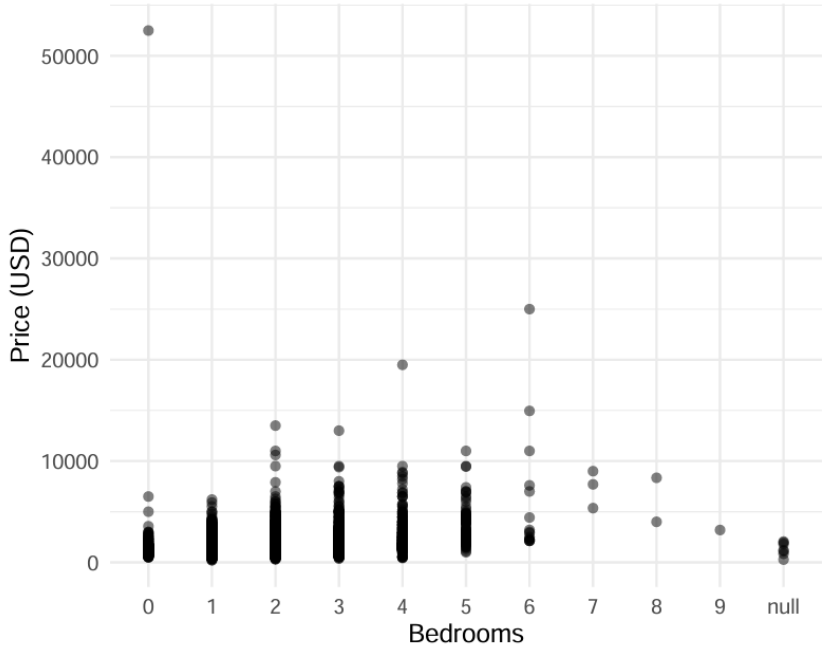


Hình: Phân phối diện tích căn hộ

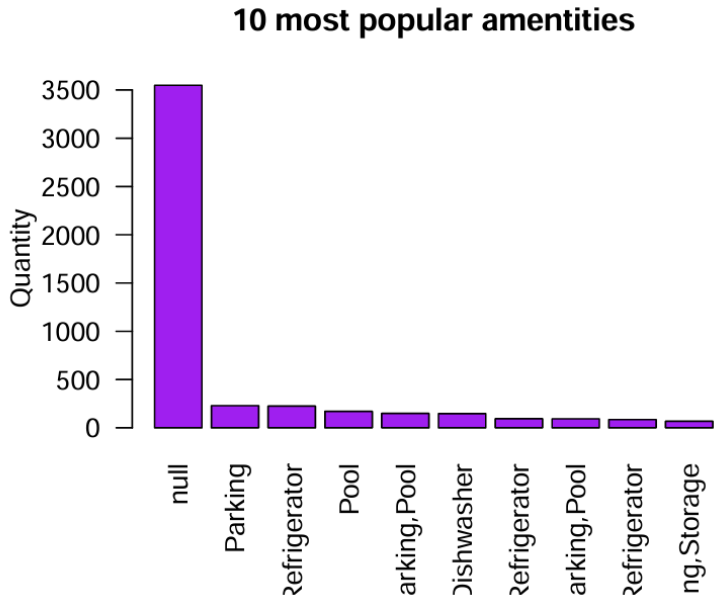


Hình: Mối quan hệ giữa số phòng tắm và giá thuê

EDA



Hình: Mối quan hệ giữa số phòng ngủ và giá thuê



Hình: 10 tiện ích phổ biến nhất

Tiền xử lý dữ liệu

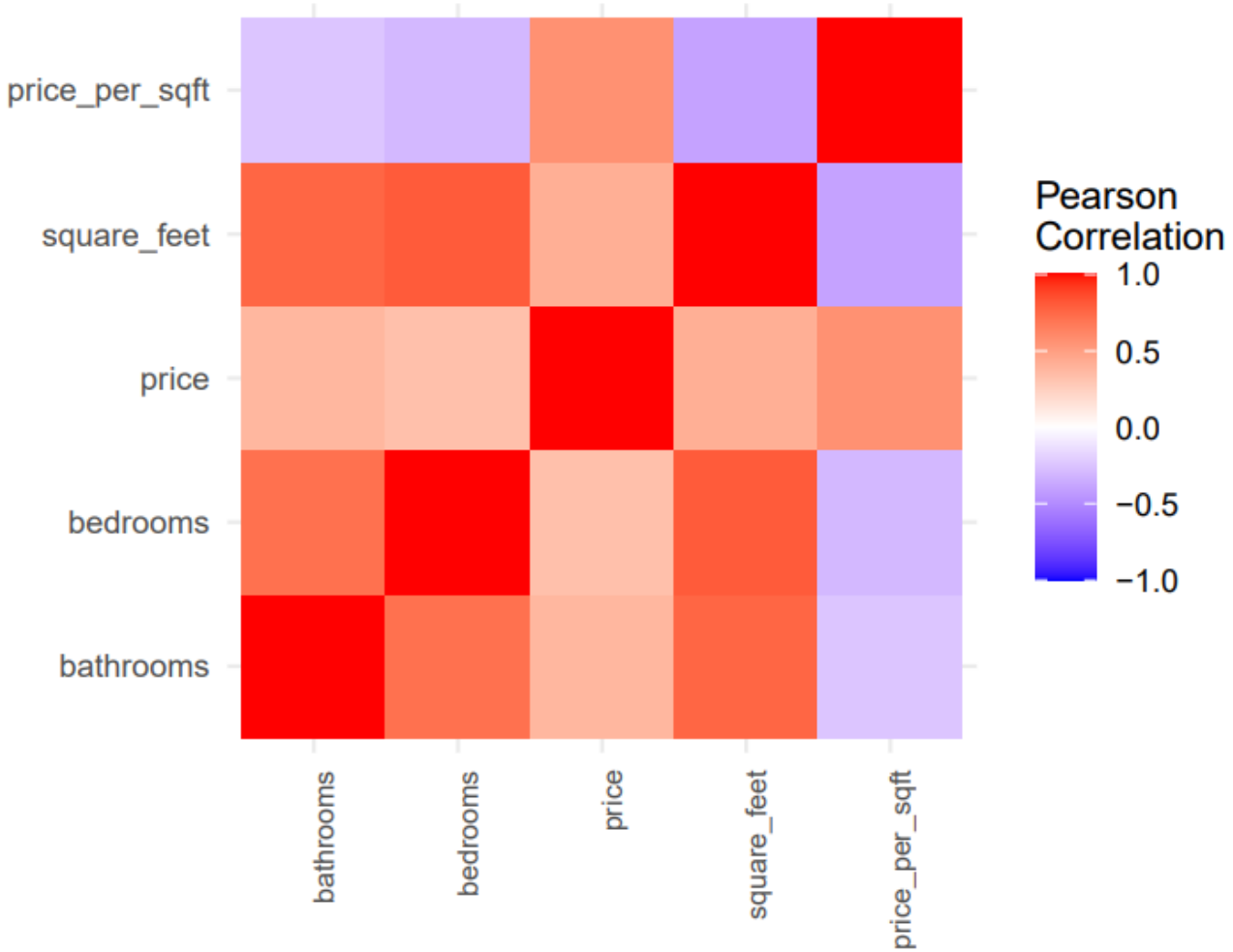
- Loại bỏ các cột không cần thiết: id, category, title, body, ... (chủ yếu là các cột chứa thông tin về căn hộ bằng text mà không liên quan đến xử lý).
- Xử lý dữ liệu thiếu: có thể là sẽ là điền dữ liệu khuyết (nếu số lượng dữ liệu khuyết là đáng kể trên tập dữ liệu) hoặc loại bỏ dữ liệu khuyết (nếu số lượng dữ liệu khuyết là không đáng kể trên tập dữ liệu).
- Xử lý cột `cityname`: Gộp các thành phố có ít hơn 10 căn hộ vào nhóm "other"
- Lọc dữ liệu: Lọc các hàng thỏa điều kiện diện tích chia cho số phòng ngủ nhỏ hơn 250, vì các căn hộ có diện tích như vậy là quá nhỏ đối với số phòng ngủ được cung cấp. Ngoài ra, tính giá trị trung bình của mỗi mét vuông từ hai cột dữ liệu là giá thuê và diện tích, bởi vì khi tham khảo giá thuê căn hộ, nhiều người sẽ quan tâm đến giá tiền trên 1 mét vuông

Tiền xử lý dữ liệu

- Loại bỏ các điểm ngoại lai: Tạo một hàm để loại bỏ ngoại lệ dựa trên mean và standard deviation của `price_per_sqft`, `square_feet` và `price` theo từng thành phố, lọc các dòng dữ liệu ngoài phạm vi mean $\pm$ 3*std. Chỉ lấy các dòng trong đó có số phòng ngủ, số giường, số phòng tắm, giá tiền lớn hơn 0
- Tạo biến giả cho cột `cityname`

Tiền xử lý dữ liệu

- Tính toán ma trận tương quan giữa các biến số



Xây dựng mô hình hồi quy đơn giản

- Trước khi tiến hành xây dựng mô hình hồi quy đơn giản, ta sẽ chia tập dữ liệu thành hai tập là huấn luyện (train_data) và kiểm thử (test_data) với tỉ lệ là 7/3.

```
## cityname_west_lafayette    3.773e-03  7.810e-02   0.048  0.96147
## cityname_west_new_york     9.639e-02  9.412e-02   1.024  0.30584
## cityname_winston_salem    -3.852e-02  9.014e-02  -0.427  0.66911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1645 on 6094 degrees of freedom
## Multiple R-squared:  0.8514, Adjusted R-squared:  0.847
## F-statistic:   193 on 181 and 6094 DF,  p-value: < 2.2e-16
```

Hình: Bảng thống kê tổng hợp của mô hình hồi quy đơn giản

Xây dựng mô hình hồi quy đơn giản

- Mô hình hồi quy tuyến tính đơn giản giải thích được khoảng 85% biến thiên của giá thuê căn hộ.
- Giá trị RSE không phải là hoàn hảo nhưng vẫn cho thấy mô hình có khả năng dự đoán tương đối chính xác.
- Vì giá trị p-value của các features bathrooms, bedrooms, square_feet và price_per_sqft rất nhỏ (đều dưới 0.001) nên các biến này có ảnh hưởng mạnh và có ý nghĩa thống kê cao đến giá căn hộ.

Xây dựng mô hình hồi quy đơn giản

```r
# Dự đoán trên tập test
predictions <- predict(lm_model, newdata = test_data)

# Đánh giá model
RMSE <- sqrt(mean((test_data$price - predictions)^2))
cat("Root Mean Squared Error (RMSE):", RMSE, "\n")
```

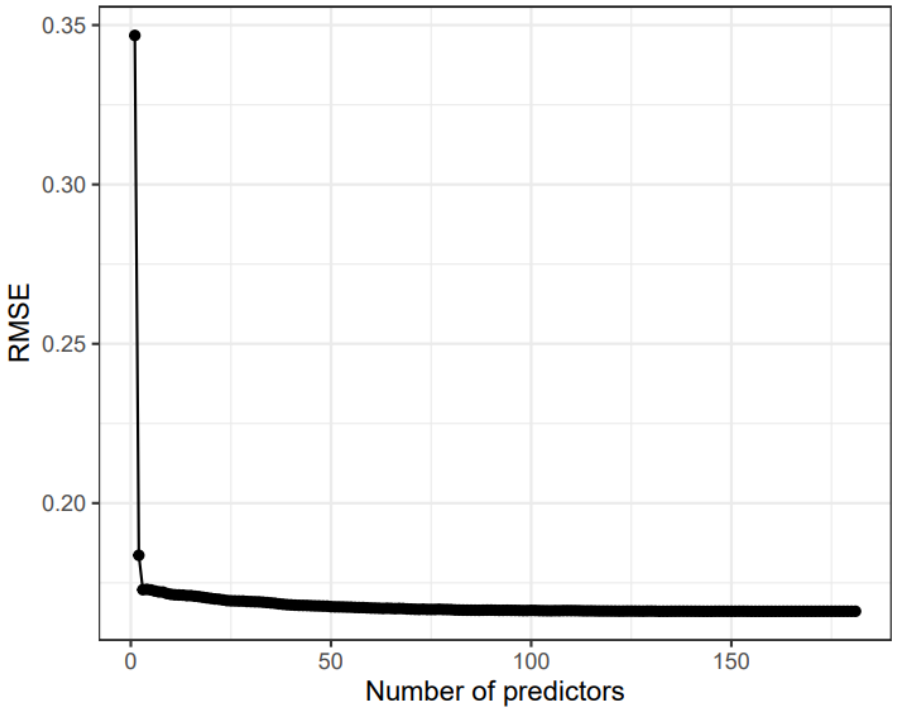
```
## Root Mean Squared Error (RMSE): 0.1663875
```

Hình: Đánh giá mô hình khi dự đoán trên tập test

Xây dựng mô hình hồi quy đơn giản

- RMSE chỉ hơi cao hơn một chút so với RSE trên tập huấn luyện (0.1645) $\rightarrow$ Mô hình có khả năng dự đoán chính xác và không bị overfitting nhiều.
- Tuy nhiên, ta sẽ thực hiện cross validation để đảm bảo tính chính xác, tin cậy và đánh giá mô hình một cách toàn diện hơn.

Hồi quy từng bước với cross validation



Hình: RMSE của 181 mô hình con được xây dựng bằng phương pháp "forward"

Sau khoảng 100 biến dự báo, đường cong RMSE không giảm nhiều nữa, thậm chí có xu hướng tăng nhẹ. Điều này gợi ý rằng việc thêm quá nhiều biến dự báo vào mô hình có thể không cải thiện đáng kể độ chính xác dự báo mà còn có thể dẫn đến hiện tượng overfitting.

Hồi quy từng bước với cross validation

- Giá trị nhỏ nhất của 10-fold cross-validated error xảy ra khi số lượng biến dự báo trong mô hình là 142. Đây sẽ là giá trị tối ưu về độ chính xác dự báo đối với tập dữ liệu này.
  $\rightarrow$ Sử dụng 142 biến dự báo quan trọng nhất để xây dựng mô hình mới.

Xây dựng mô hình từ kết quả của cross validation

```
## cityname_west_lafayette   -1.305e-03  2.678e-02  -0.049 0.961156
## cityname_west_new_york     9.100e-02  5.868e-02   1.551 0.120976
## cityname_winston_salem    -4.356e-02  5.228e-02  -0.833 0.404746
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1642 on 6132 degrees of freedom
## Multiple R-squared:  0.8511, Adjusted R-squared:  0.8476
## F-statistic: 245.1 on 143 and 6132 DF,  p-value: < 2.2e-16
```

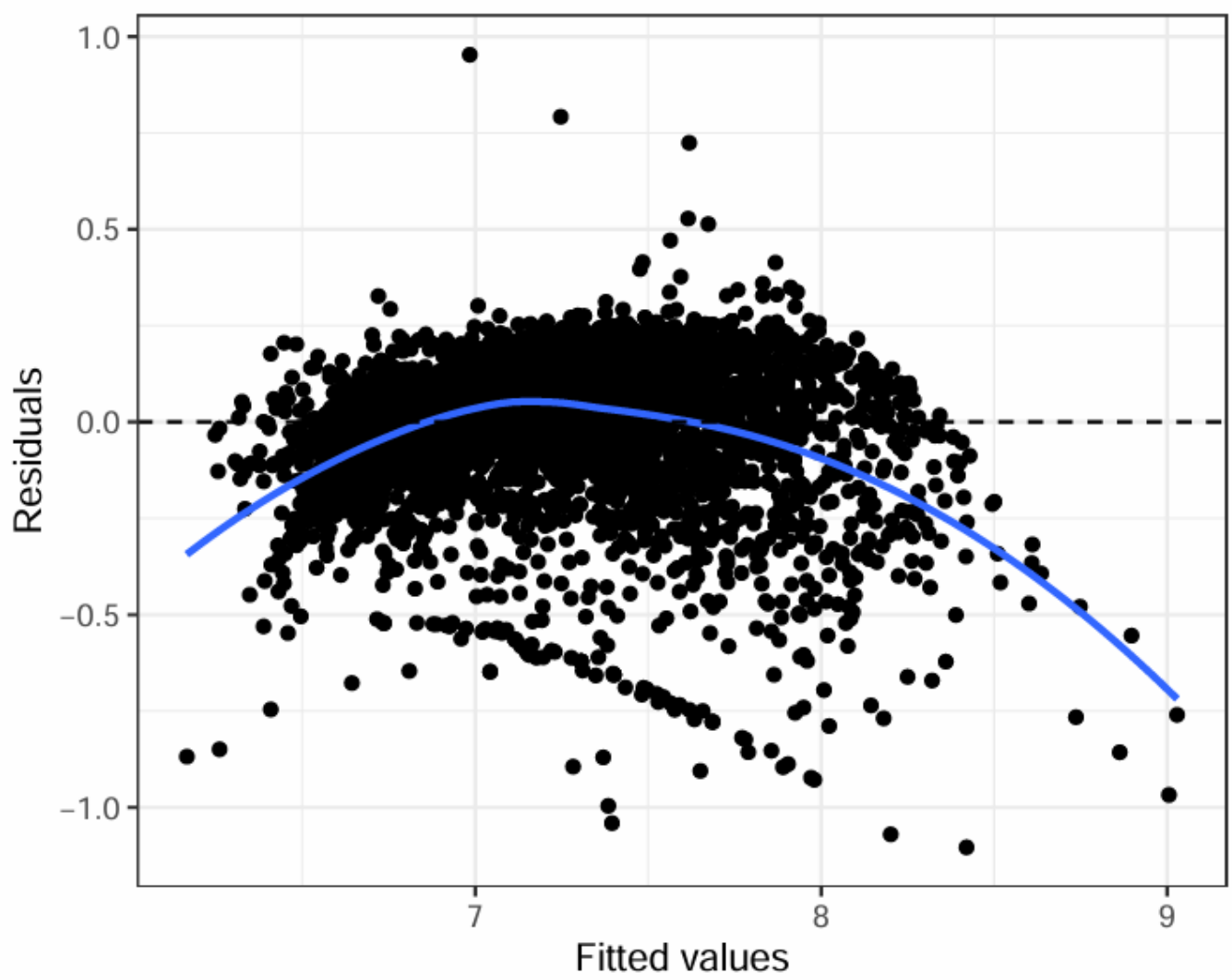
Hình: Bảng thống kê tổng hợp của mô hình được xây dựng từ kết quả của cross validation

```
## Root Mean Squared Error (RMSE): 0.1656222
```

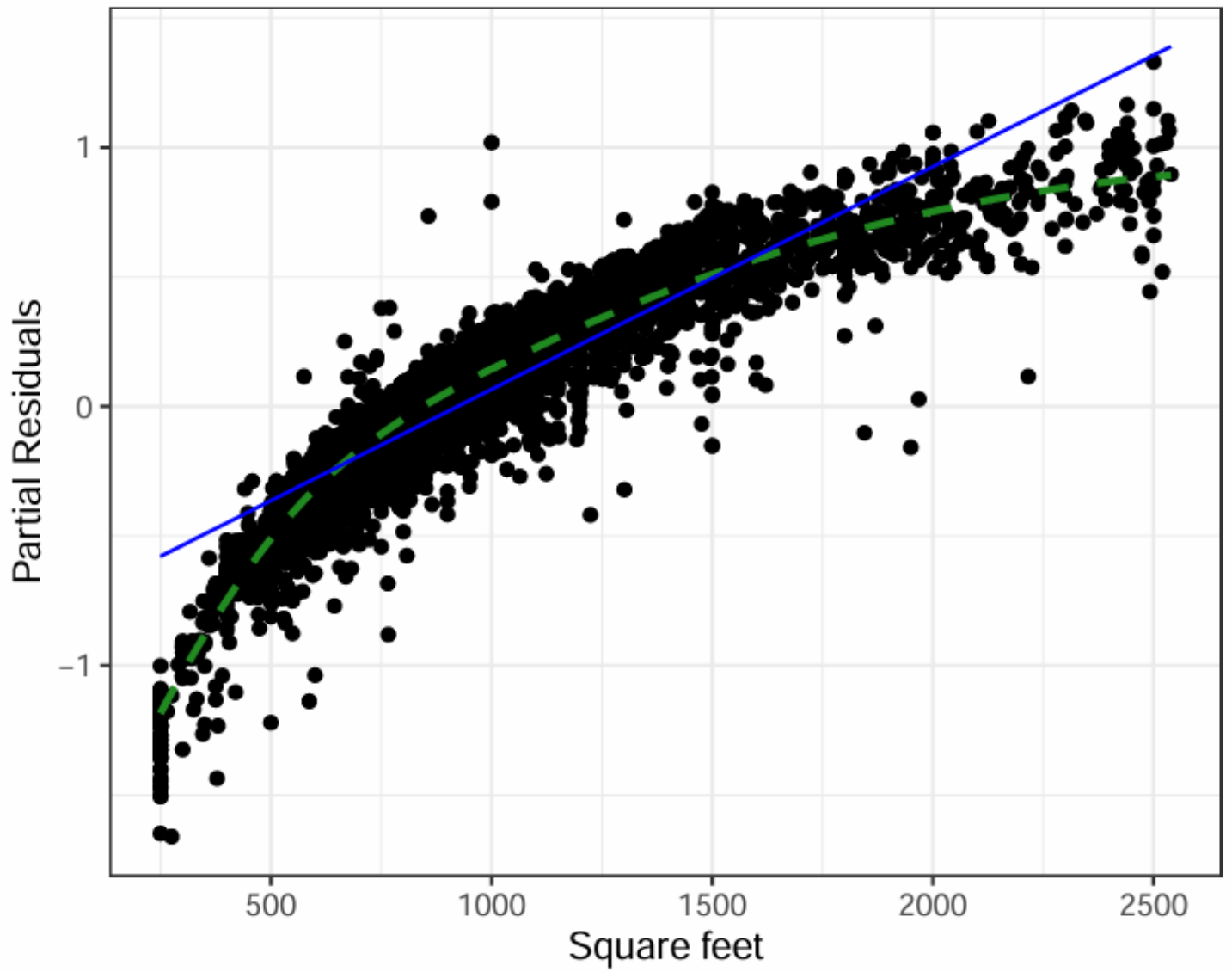
Xây dựng mô hình từ kết quả của cross validation

- Mô hình hồi quy tuyến tính với các biến tối ưu có độ chính xác cao, giải thích được khoảng 85% biến thiên của giá thuê căn hộ.
- Giá trị F-statistic lớn và p-value rất nhỏ cho thấy mô hình có ý nghĩa thống kê cao.
- Residual standard error (RSE) là 0.1642, cho thấy sai số dự đoán trung bình rất thấp.
- RMSE thấp trên cả tập huấn luyện và tập kiểm tra $\rightarrow$ Mô hình hồi quy tuyến tính với 142 biến dự báo tối ưu đã được xây dựng khá tốt và có khả năng dự đoán chính xác trên dữ liệu mới.

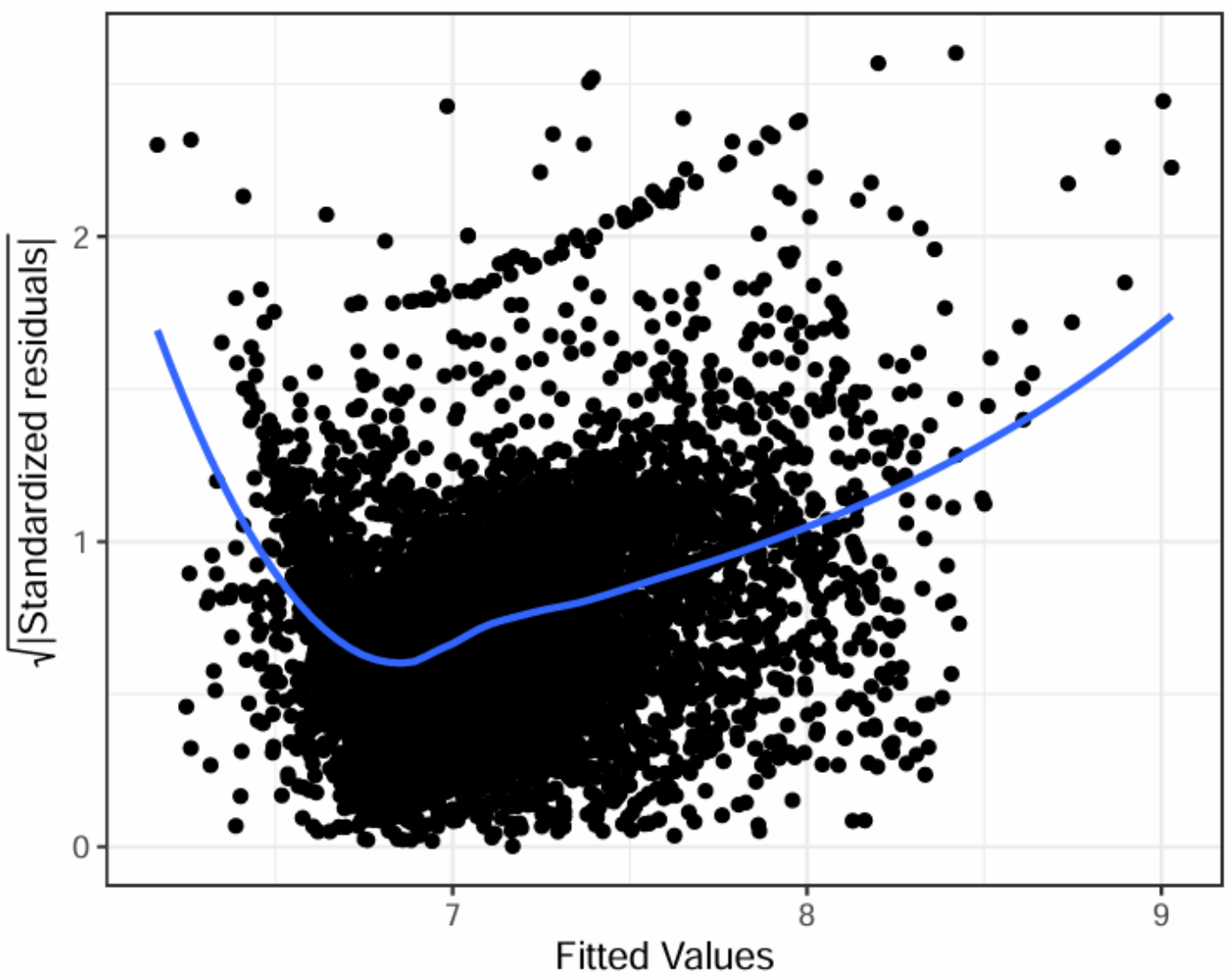
Biểu đồ thặng dư của mô hình



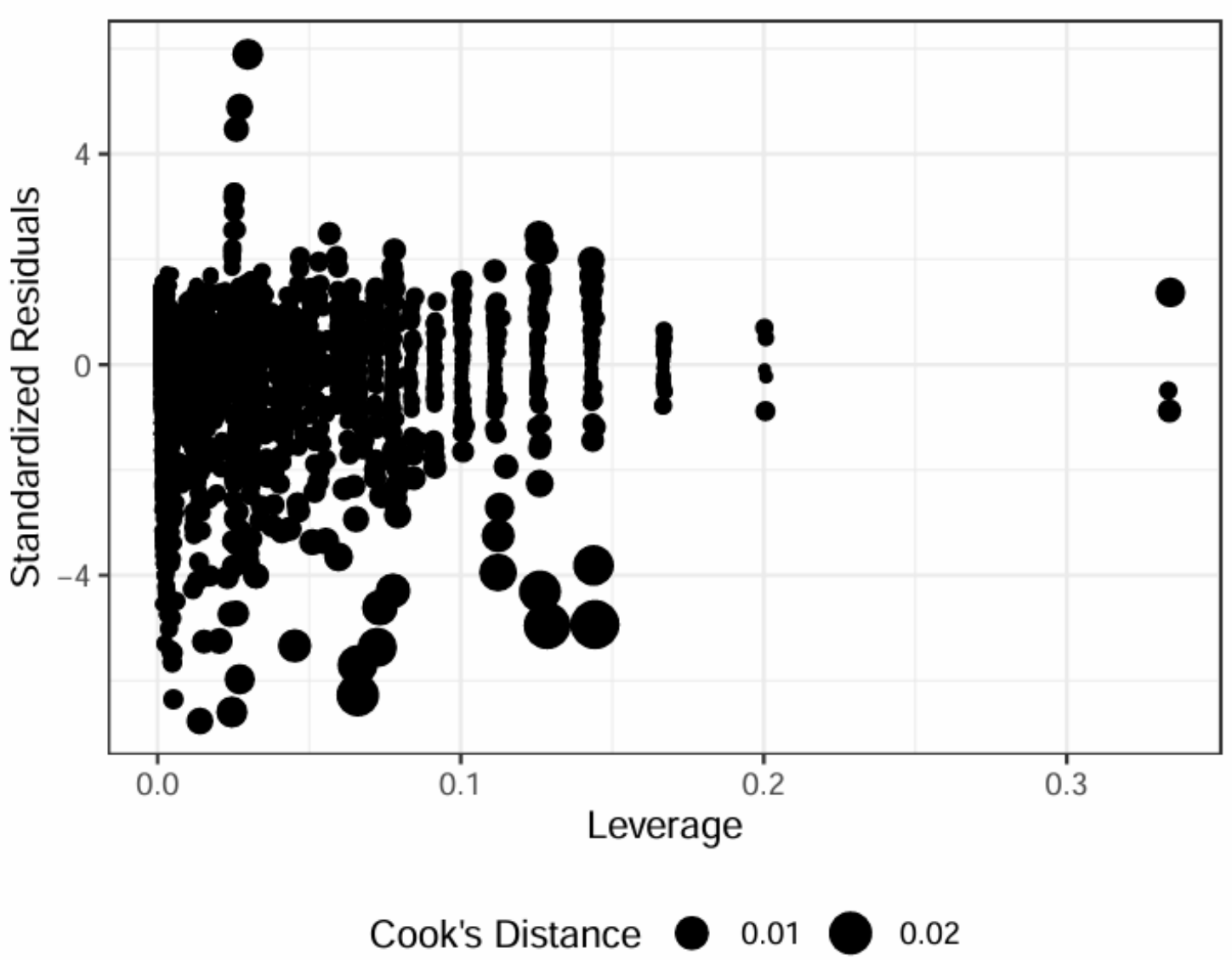
Kiểm tra tính tuyến tính từng phần



Kiểm tra tính đồng nhất phương sai



Kiểm tra điểm ngoại lai trong mô hình



Xây dựng mô hình sau khi loại bỏ outliers

Dựa vào kết quả từ biểu đồ cooks, ta sẽ thử loại bỏ các điểm ngoại lai từ `train_data` để xây dựng mô hình dựa trên dữ liệu mới này và đánh giá hiệu suất.

```
## # A tibble: 221 x 5
##    id_point rstand   hats  cooks sales
##       <int>  <dbl>  <dbl>  <dbl> <dbl>
##  1     2545  -4.94 0.144  0.0285  6.91
##  2     4446  -4.96 0.129  0.0252  8.27
##  3     5897  -6.28 0.0661 0.0193  6.39
##  4     4445  -4.31 0.126  0.0186  7.59
##  5     2401  -3.81 0.144  0.0170  6.80
##  6     4444  -5.70 0.0660 0.0160  6.75
##  7     4053  -5.37 0.0725 0.0156  5.41
##  8     2730  -3.95 0.112  0.0137  6.75
##  9     2591  -4.62 0.0734 0.0117  6.84
## 10     2808  -4.29 0.0777 0.0108  5.97
## # i 211 more rows
```

Xây dựng mô hình sau khi loại bỏ outliers

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1417 on 5912 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8825
## F-statistic: 321.2 on 142 and 5912 DF,  p-value: < 2.2e-16
```

Hình: Bảng thống kê tổng hợp của mô hình hồi quy sau khi loại bỏ outliers

- Residual standard error giảm từ 0.1427 xuống 0.1417, chỉ ra mô hình dự báo tốt hơn.
- Multiple R-squared tăng từ 0.8797 lên 0.8853, tức mô hình giải thích được nhiều phần biến thiên của biến phụ thuộc hơn.
- Adjusted R-squared tăng từ 0.8768 lên 0.8825, chỉ ra mô hình có khả năng tổng quát hóa tốt hơn.
- F-statistic cũng tăng, chứng tỏ mô hình có ý nghĩa thống kê tốt hơn.

Xây dựng mô hình hồi quy bậc 2

Bên cạnh đó, thông qua biểu đồ thặng dư từng phần cho biến `square_feet`, ước lượng tuyến tính là không phù hợp với dữ liệu. Vậy ta có thể dùng tập dữ liệu này mở rộng thành phần `square_feet` lên bậc 2 để ước lượng mô hình.

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1019 on 5873 degrees of freedom
## Multiple R-squared:  0.9411, Adjusted R-squared:  0.9393
## F-statistic: 518.4 on 181 and 5873 DF,  p-value: < 2.2e-16
```

Hình: Bảng thống kê tổng hợp của mô hình hồi quy sau bậc 2 cho `square_feet`

Xây dựng mô hình hồi quy bậc 2

- Độ chính xác của mô hình rất cao với Adjusted R-squared là 0.9393, giải thích được 93.93% biến thiên của biến phụ thuộc (`price`).
- Các biến dự báo có mức ý nghĩa thống kê cao, với hầu hết p-values $<$ 0.001, cho thấy đóng góp đáng kể vào việc dự báo giá.
- Hiệu suất dự báo của mô hình rất tốt với Residual standard error chỉ 0.1019, chứng tỏ độ lệch chuẩn của phần dư rất thấp.

Nhìn chung, việc thêm bậc 2 của biến `square_feet` đã giúp cải thiện đáng kể chất lượng của mô hình hồi quy tuyến tính. Mô hình hồi quy tuyến tính bậc 2 với biến `square_feet` trên là mô hình phù hợp nhất để dự báo giá thuê căn hộ.

- Mô hình giúp hiểu rõ hơn mối quan hệ giữa diện tích căn hộ và giá thuê. Mô hình cho thấy giá thuê tăng không tuyến tính theo diện tích, với tốc độ tăng nhanh hơn ở các căn hộ có diện tích lớn hơn.
- Từ đó, có thể đề xuất các kích cỡ căn hộ phù hợp với nhu cầu sử dụng và khả năng chi trả của người thuê. Ví dụ, với người thuê cá nhân, căn hộ $50 - 70m^2$ có thể là lựa chọn hợp lý. Còn với gia đình đông người, căn hộ $80 - 100m^2$ sẽ phù hợp hơn.
- Giúp các công ty môi giới có cơ sở khoa học hơn trong việc định giá và tư vấn cho khách hàng, từ đó nâng cao chất lượng dịch vụ.

Kết quả đạt được

```
> specific_predictions <- predict(lm_model_poly, newdata = test_data[1:5,])
> specific_predictions <- exp(specific_predictions)
> print(specific_predictions)
     2159      2593      3043      3509      3676
 799.4694  854.1538 1314.2307  913.8349  992.4659
> actual_prices <- exp(test_data$price[1:5])
> print(actual_prices)
[1]  840  905 1424  965 1065
```

Hình: Giá trị dự đoán của mô hình

Mô hình đưa ra giá trị dự đoán tương tối fit với giá trị thực tế, điều này cho thấy mô hình dự đoán có thể "dùng được" trong khả năng tiếp cận đến các giá thuê nhà.

Kết quả đạt được

- Vị trí địa lý, diện tích, số lượng phòng và tiện ích là những yếu tố chính ảnh hưởng đến giá thuê nhà. Căn hộ tại các thành phố lớn, diện tích rộng, nhiều phòng và nhiều tiện ích sẽ có giá thuê cao hơn.
- Người thuê nhà cần cân nhắc kỹ các yếu tố như vị trí, diện tích, tiện ích và nghiên cứu kỹ thị trường để chọn căn hộ phù hợp với nhu cầu và tài chính.
- Các công ty môi giới cần tư vấn khách hàng dựa trên phân tích dữ liệu chi tiết về thị trường và các yếu tố ảnh hưởng đến giá thuê, giúp khách hàng đưa ra quyết định tốt nhất, đồng thời theo dõi sát sao diễn biến thị trường.

Thank you for your attention!