**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

# PHÂN TÍCH GIÁ CHUNG CƯ TRÊN ĐỊA BÀN THÀNH PHỐ HỒ CHÍ MINH



| **Nhóm 06** | | | |
|---|---|---|---|
| Sinh viên thực hiện: | | | |
| STT | Họ tên | MSSV | Ngành |
| 1 | Nguyễn Tuệ Minh | 21521140 | CNCL |
| 2 | Nguyễn Phúc Khang | 21522194 | CNCL |
| 3 | Đỗ Nguyễn Anh Khoa | 21522219 | CNCL |



**TP. HỒ CHÍ MINH – 12/2024**

# 1. GIỚI THIỆU

Đề tài tập trung vào việc phân tích dữ liệu, áp dụng các mô hình học máy và học sâu nhằm dự đoán giá chung cư tại thành phố Hồ Chí Minh, dựa trên các thông tin được thu thập từ năm 2020. Để thực hiện điều này, nhóm đã tiến hành nghiên cứu rõ từng thuộc tính trong bộ dữ liệu gốc nhằm hiểu rõ dữ liệu trước khi phân tích. Sau đó tiến hành phân tích thăm dò dữ liệu, tiền xử lý, làm sạch dữ liệu bằng cách xử lý các giá trị khuyết, định dạng dữ liệu, chuẩn hoá dữ liệu, phát hiện và xử lý ngoại lệ nhằm đảm bảo tính chính xác của bộ dữ liệu. Sau khi xử lý, các thuộc tính quan trọng được lựa chọn để xây dựng các mô hình học máy. Nhóm áp dụng 3 mô hình dự đoán là Neural Network, Linear Regression và mô hình Ensemble sử dụng kỹ thuật Gradient Boosting Machines. Sau quá trình tinh chỉnh, đánh giá, mô hình có hiệu suất dự đoán tối ưu nhất sẽ được lựa chọn. Mục tiêu của đề tài là ứng dụng những kỹ thuật phân tích dữ liệu được học nhằn khám phá những hiểu biết sâu sắc từ dữ liệu thông qua bài toán dự đoán giá chung cư từ đó có thể hỗ trợ việc ra quyết định trong các dự án thực tế sau này. Bên cạnh đó, nhóm xây dựng mô hình học máy có khả năng dự đoán chính xác giá cho thuê chung cư dựa trên các thuộc tính quan trọng đã được chọn lọc.

Trong đề tài này, chúng tôi đã sử dụng bộ dữ liệu phân tích được tham khảo tại nguồn [1]. Trong đó, chúng tôi đã tiến hành tiền xử lý, trực quan hoá lại bộ dữ liệu, xây dựng mô hình học máy sử dụng các thuật toán khác với dự án mẫu từ đó có sự so sánh và đánh giá khách quan hơn nhằm đóng góp và cải thiện cho bài toán ban đầu.

# 2. MÔ TẢ BỘ DỮ LỆU (BẮT BUỘC)

a. Tóm tắt bộ dữ liệu

Bộ dữ liệu được thu thập từ trang Chợ tốt[2] về thông tin các căn chung cư được bán trên địa bàn thành phố Hồ Chí Minh năm 2020.

Bộ dữ liệu phân tích được tham khảo tại nguồn [1].

Bộ dữ liệu gồm 24.949 mẫu, với 18 thuộc tính.

b. Liệt kê và giải thích các Features

| STT | Feature | Kiểu dữ liệu |
|---|---|---|
| 1 | DiaChi | Object |
| 2 | Quan | Object |
| 3 | TinhTrangBDS | Object |
| 4 | DienTich | Float64 |
| 5 | Phongngu | Float64 |
| 6 | TenPhanKhu | Object |
| 7 | SoTang | Float64 |
| 8 | PhongTam | Float64 |

| | | |
|---|---|---|
| 9 | Loai | Object |
| 10 | GiayTo | Object |
| 11 | MaCanHo | Object |
| 12 | TinhTrangNoiThat | Object |
| 13 | HuongCuaChinh | Object |
| 14 | HuongBanCong | Object |
| 15 | DacDiem | Object |
| 16 | Gia | Float64 |
| 17 | USD | Float64 |
| 18 | Log_price | Float64 |

Giải thích các Features

1. **DiaChi**: Địa chỉ của chung cư (tại TpHCM). Đây là chuỗi ký tự đại diện cho vị trí chính xác của chung cư.
2. **Quan**: Quận nơi chung cư tọa lạc. Đây là chuỗi ký tự, giúp phân loại chung cư theo khu vực hành chính cấp quận.
3. **TinhTrangBDS**: Tình trạng của bất động sản. Đây là chuỗi ký tự, có thể mô tả chung còn mới, cũ hoặc đang được sửa chữa.
4. **DienTich**: Diện tích của chung cư, đo bằng mét vuông ($m^2$). Đây là giá trị số thập phân (float).
5. **Phongngu**: Số lượng phòng ngủ trong của từng chung cư. Đây là giá trị số thập phân.
6. **TenPhanKhu**: Tên phân khu nơi bất động sản thuộc về. Đây là chuỗi ký tự giúp xác định khu vực nhỏ hơn trong một quận.
7. **SoTang**: Số tầng của toàn bộ chung cư đó. Đây là giá trị số thập phân.
8. **PhongTam**: Số lượng phòng tắm của từng chung cư. Đây là giá trị số thập phân.
9. **Loai**: Loại bất động sản, ví dụ như căn hộ, biệt thự, nhà phố. Được biểu diễn là chuỗi ký tự.
10. **GiayTo**: Loại giấy tờ pháp lý của bất động sản, như sổ đỏ, sổ hồng. Đây là chuỗi ký tự.
11. **MaCanHo**: Mã căn hộ (nếu có), giúp xác định riêng biệt từng căn hộ trong một khu chung cư. Đây là chuỗi ký tự, nhưng có nhiều giá trị thiếu.
12. **TinhTrangNoiThat**: Tình trạng nội thất của bất động sản, ví dụ như đã hoàn thiện hay chưa, hoặc còn trống. Đây là chuỗi ký tự và có nhiều giá trị thiếu.

13. **HuongCuaChinh**: Hướng cửa chính của bất động sản, ví dụ như Đông, Tây, Nam, Bắc. Được biểu diễn bằng chuỗi kí tự
14. **HuongBanCong**: Hướng ban công (nếu có), mô tả hướng của ban công bất động sản. Đây là chuỗi ký tự.
15. **DacDiem**: Các đặc điểm nổi bật của bất động sản, ví dụ như có hồ bơi, sân vườn. Đây là chuỗi ký tự, có nhiều giá trị thiếu.
16. **Gia**: Giá trị của bất động sản tính theo Việt Nam đồng. Đây là giá trị số thập phân.
17. **USD**: Giá trị của bất động sản tính theo USD. Đây là giá trị số thập phân.
18. **log_price**: Giá trị logarit của giá bất động sản, dùng để làm mịn dữ liệu giá. Đây là giá trị số thập phân.

c. Phương pháp xử lý giá trị khuyết cho các Features khuyết

Vì bộ dữu liệu được crawl từ trang Chợ tốt nên không thể tránh khỏi việc bộ data có các thiếu soát và mất dữ liệu. Dưới đây là thống kê các sample bị thiếu dữ liệu và phần trăm của chúng theo từng feature:

| | Null Count | Null Percentage (%) |
|---|---|---|
| DiaChi | 0 | 0.000000 |
| Quan | 0 | 0.000000 |
| TinhTrangBDS | 2 | 0.008327 |
| DienTich | 0 | 0.000000 |
| Phongngu | 0 | 0.000000 |
| TenPhanKhu | 17154 | 71.418460 |
| SoTang | 17490 | 72.817353 |
| PhongTam | 0 | 0.000000 |
| Loai | 0 | 0.000000 |
| GiayTo | 5733 | 23.868604 |
| MaCanHo | 20738 | 86.339981 |
| TinhTrangNoiThat | 11599 | 48.290936 |
| HuongCuaChinh | 14852 | 61.834381 |
| HuongBanCong | 15517 | 64.603023 |
| DacDiem | 18588 | 77.388734 |
| Gia | 0 | 0.000000 |
| USD | 0 | 0.000000 |
| log_price | 0 | 0.000000 |

Tùy và cột thuộc tính feature thì nhóm chúng em sẽ có các các sử lý giá trị null khác nhau.

- Xóa các sample bị null : Đối với những thuộc tính có sample bị null quá ít, việc tối ưu nhất là xóa đi các sample đó tránh việc ảnh hưởng tới bộ dữ liệu. Chúng em tiến hành thực hiện phương pháp đó đối với feature TinhTrangBDS chỉ có 2 sample có giá trị null, nếu xóa sẽ không gây ra ảnh hưởng.

- Xóa feature : Đối với những feature không mang lại quá nhiều ảnh hưởng tới việc quyết định giá nhà, feature không có tính đóng góp cho việc phân tích như là :
    - TenPhanKhu : Tên phân khu sẽ tùy vào căn trung cư đó thuộc vào chung cư nào sẽ có cách đặt tên phân và phân loại khác nhau. Trong bộ dữ liệu có tới 900 giá trị khác nhau, sẽ không đồng nhất cách gọi giữa các căn chung cư. Dẫn đến việc không thể sử dụng để phân tích giá nhà
    - SoTang : Tương tự, số tầng của từng căn hộ khác nhau tùy vuộc vào từng dự án chung cư. Vì số tần không đống góp quan trọng trong phân tích giá giữa các căn chúng cư với nhay, chúng em quyết định bỏ feature này
    - DacDiem : Ở đây sau khi nhóm chúng em phân tích và thấy rằng, thuộc tính DacDiem chỉ có một giá trị là "căn góc" còn lại là giá trị null, và số lượng data bị null chiếm rất nhiều lên tới 77%
    - MatCanHo : Không như mua nhà, mặt căng chung cư không ảnh hưởng quá nhiều đối với giá nhà, theo bộ dữ liệu cho thấy có tới 86% sample không nêu ra mặt căn hộ khi bán, cho thấy đây không phải yếu tố quá ảnh hưởng tới giá nhà
    - HuongCuaChinh : tương tự với mặt căn hộ, hướng cửa chính cũng vậy. Khi bộ dữ liệu null tới 62%
    - HuongBanCon : tương tự với mặt căn hộ, hướng ban công cũng vậy. Khi bộ dữ liệu null tới 65%
- Sử dụng thuật toán Random Forest để điền các giá trị null : Đối với các feature còn lại (GiayTo, TinhTrangNoiThat) là những đặc trung quan trọng trong việc ảnh hưởng đến việc quyết định giá nhà, nên nhóm chúng em quyết định giữ lại và điền giá trị cho chúng. Do các feature này tất cả là dạng nhãn, nên nhóm chúng em đã tiến hành đánh giá và tìm ra thuật toán phù hợp để điền. Trong các thuật toán máy học mà nhóm chúng em đã được học thì thuật toán RandomForest đã cho độ chính xác cao nhất trên 90% ở các feature.
    - Đầu tiên nhóm chúng em sẽ lọc ra các sample và các thuộc tính không có giá trị bị khuyết để chuẩn bị bộ dữ liệu bộ dữ liệu để huấn luyện mô hình và đánh giá kết quả
    - Sau đó tiến hành lần lực với từng các feature bị khuyết (GiayTo, TinhTrangNoiThat)
    - Sau khi đã điền xong chúng em tiến hành gộp vô lại bộ dữ liệu để tạo thành bộ dữ liệu hoàn chỉnh.

# 3. PHƯƠNG PHÁP PHÂN TÍCH

Về phương pháp phân tích, nhóm đã tiến hành và thực hiện theo quy trình:

Import và thăm dò dữ liệu → Xử lý và chọn đặc trưng → Xây dựng mô hình Machine Learning và Deep Learning → Đánh giá hiệu quả mô hình.

## 3.1. Import và thăm dò dữ liệu

**Exploratory Data Analysis (EDA)** hay thăm dò dữ liệu là bước đầu tiên trong quá trình phân tích dữ liệu, tập trung vào việc khám phá, tóm tắt và trực quan hóa các đặc điểm chính của tập dữ liệu. Mục tiêu của EDA là hiểu rõ hơn về cấu trúc, mẫu, và các vấn đề tiềm tàng trong dữ liệu, giúp đưa ra quyết định hoặc chuẩn bị tốt hơn cho các bước phân tích hoặc mô hình hóa sau này. Trong đó nhóm tiến hành thực hiện các bước sau:

- Handling Missing Data: Xử lý dữ liệu bị thiếu nhằm loại bỏ hoặc thay thế các giá trị còn trống trong tập dữ liệu, đảm bảo dữ liệu đầy đủ và chất lượng cao để phục vụ phân tích. Nhằm giảm thiểu ảnh hưởng của giá trị thiếu lên kết quả phân tích, bảo toàn thông tin bằng cách sử dụng các phương pháp thay thế phù hợp như giá trị trung bình, trung vị, hoặc các kỹ thuật tiên tiến như MICE (Multiple Imputation by Chained Equations) và tránh lỗi khi áp dụng các thuật toán học máy vốn không hoạt động với dữ liệu thiếu.
- Data Normalization: Chuẩn hóa dữ liệu nhằm đưa tất cả các đặc trưng về cùng một thang đo để loại bỏ sự khác biệt về đơn vị hoặc phạm vi giá trị giữa các biến. Đảm bảo rằng các thuật toán nhạy cảm với thang đo như KNN, SVM, và Gradient Descent hoạt động ổn định hơn. Tăng khả năng phân tích khi dữ liệu có sự chênh lệch lớn về thang đo. Ngoài ra, còn giúp dữ liệu trở nên dễ trực quan hóa hơn, chẳng hạn như khi so sánh các biến trong cùng một biểu đồ.
- Data Visualization: Trực quan hóa dữ liệu nhằm biểu diễn thông tin từ tập dữ liệu dưới dạng biểu đồ, đồ thị để dễ dàng phát hiện các mẫu, xu hướng và mối quan hệ. Giúp xác định mối quan hệ giữa các biến và phát hiện các mẫu hoặc xu hướng quan trọng. Đặc biệt, dễ dàng nhận biết các giá trị ngoại lệ hoặc sự phân bố dữ liệu. Cuối cùng là truyền tải thông tin một cách trực quan, dễ hiểu, hỗ trợ thuyết phục người khác hoặc ra quyết định dựa trên dữ liệu.

## 3.2. Xử lý và chọn đặc trưng

**Feature Engineering** là quá trình tạo, biến đổi và lựa chọn các đặc trưng (features) trong tập dữ liệu để tăng hiệu suất của các mô hình học máy. Đây là một bước quan trọng trong phân tích và xây dựng mô hình, giúp tập trung vào những thông tin có giá trị nhất, đồng thời cải thiện độ chính xác và hiệu quả của mô hình. Các bước chính trong Feature Engineering của nhóm bao gồm:

- Feature Mapping: Feature Mapping là quá trình biến đổi dữ liệu hiện có thành các dạng phù hợp hơn bằng cách áp dụng các phép biến đổi như logarithm, polynomial, hoặc các hàm phi tuyến khác. Công dụng nhằm giúp xử lý dữ liệu phi tuyến tính, làm nổi bật các mối quan hệ phức tạp giữa biến đầu vào và biến mục tiêu. Giảm tác động của các giá trị cực lớn hoặc cực nhỏ bằng cách áp dụng log-transform. Tạo ra

các đặc trưng mới (e.g., tương tác giữa các biến) giúp mô hình hiểu sâu hơn về dữ liệu.

- Correlation: Correlation đo lường mức độ liên kết giữa các biến, thường sử dụng các chỉ số như hệ số tương quan Pearson, Spearman hoặc Kendall.
- Feature Importance: Tính toán mức độ quan trọng của từng đặc trưng dựa trên đóng góp của chúng vào hiệu suất mô hình. Các thuật toán như XGBoost, Random Forest cung cấp thước đo này một cách tự động. Là cơ sở cho bước Feature Selection (chọn lọc đặc trưng).
- Feature Selection: Loại bỏ các đặc trưng ít quan trọng hoặc không liên quan để giảm kích thước dữ liệu và tăng hiệu suất mô hình. Có thể sử dụng các phương pháp như Recursive Feature Elimination (RFE) hoặc SelectKBest nhằm giảm thời gian huấn luyện và độ phức tạp của mô hình bằng cách giữ lại các đặc trưng quan trọng nhất. Tăng độ chính xác và hiệu suất mô hình nhờ loại bỏ nhiễu từ các đặc trưng không cần thiết.

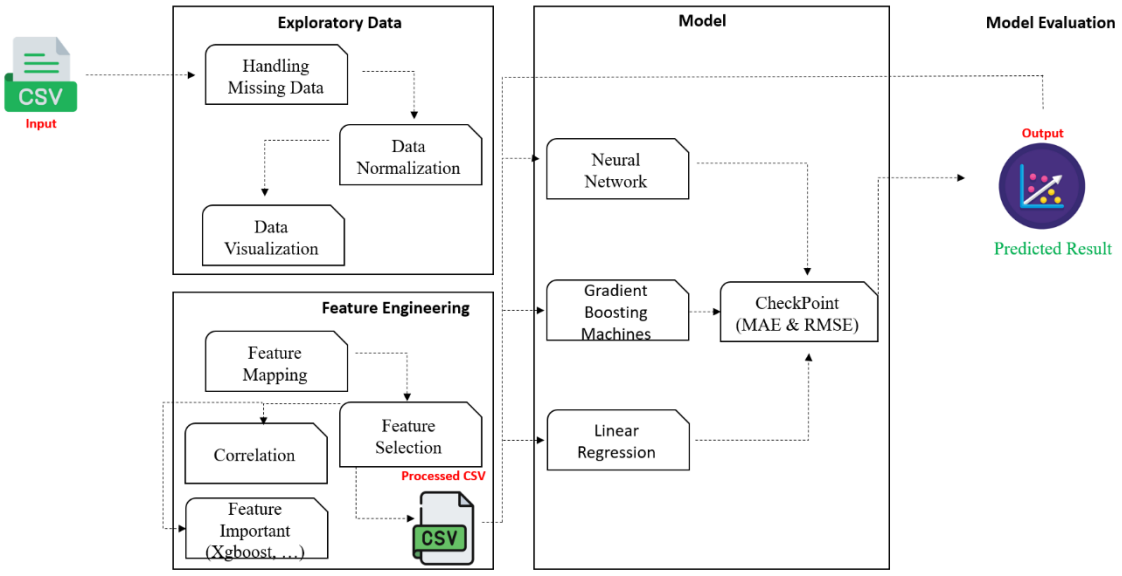
### 3.3. Xây dựng mô hình Machine Learning & Deep Learning

**Model Building** là giai đoạn quan trọng trong quy trình phân tích dữ liệu, nơi các thuật toán được áp dụng để xây dựng mô hình dự đoán hoặc phân loại. Việc lựa chọn mô hình phụ thuộc vào tính chất dữ liệu, mục tiêu phân tích, và độ phức tạp yêu cầu. Sau đây là 3 mô hình nhóm chọn và thực hiện:

- Linear Regression: Linear Regression là một phương pháp hồi quy tuyến tính nhằm mô hình hóa mối quan hệ giữa biến mục tiêu và một hoặc nhiều biến độc lập bằng cách tối ưu hóa đường thẳng phù hợp nhất. Được dùng để phân tích và dự đoán mối quan hệ tuyến tính giữa các biến, dễ dàng giải thích nhờ các hệ số hồi quy thể hiện mức độ ảnh hưởng của từng biến độc lập lên biến mục tiêu.
- Gradient Boosting Machines (GBMs): GBMs, như XGBoost, LightGBM, hay CatBoost, là các thuật toán tăng cường độ dốc, xây dựng mô hình bằng cách kết hợp nhiều cây quyết định nhỏ (weak learners) để cải thiện độ chính xác dự đoán. Hiệu suất cao trên dữ liệu phức tạp, cả bài toán phân loại và hồi quy, được tối ưu hóa để huấn luyện nhanh và giảm thiểu lỗi khi có nhiều dữ liệu lớn.
- Neural Network: Neural Network (Mạng nơ-ron nhân tạo) là một phương pháp mô phỏng hoạt động của não bộ con người, sử dụng các tầng (layers) của neural để học từ dữ liệu. Các mô hình phức tạp hơn như mạng sâu (Deep Neural Networks) có thể học các mẫu phi tuyến tính rất phức tạp.

4.Đánh giá hiệu quả mô hình

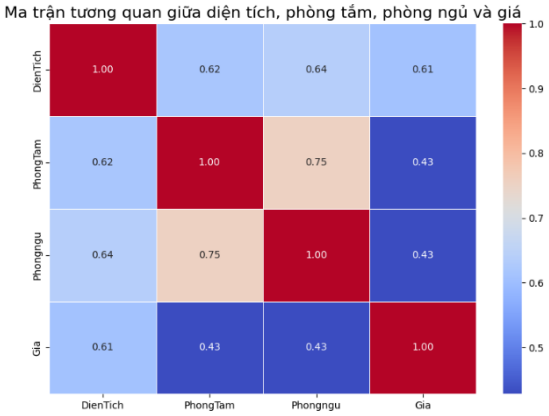
Đánh giá hiệu quả mô hình là bước cuối cùng nhưng không kém phần quan trọng trong quy trình xây dựng và phân tích mô hình. Mục tiêu là đo lường mức độ chính xác và hiệu quả của mô hình trên tập dữ liệu kiểm tra, từ đó đánh giá khả năng tổng quát hóa (generalization) của mô hình với dữ liệu mới.



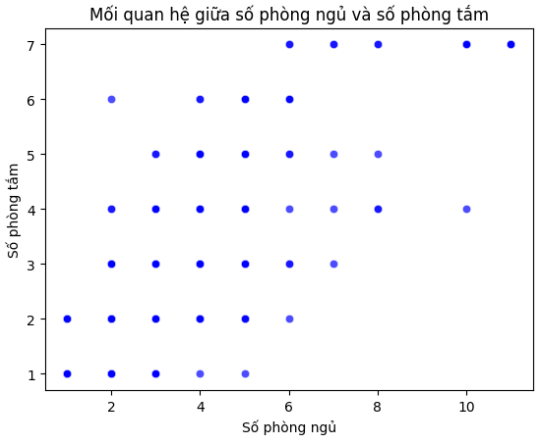

Hình 1: Phương pháp phân tích dữ liệu

## 4. PHÂN TÍCH THĂM DÒ/SƠ BỘ (gợi ý)

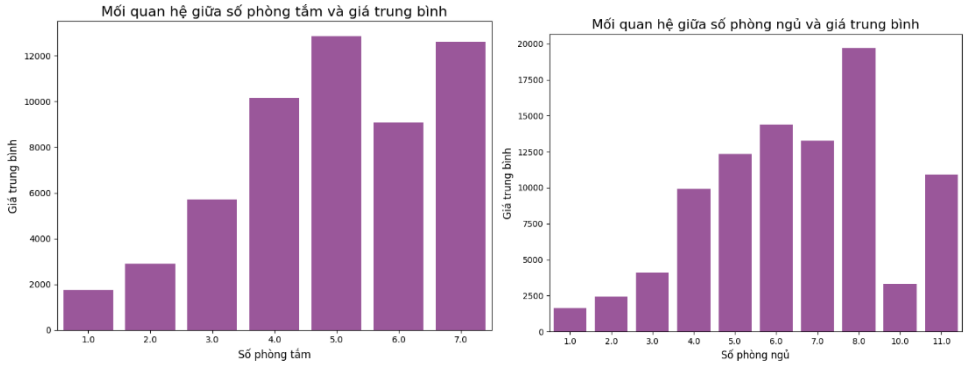
Tiến hành thăm dò sơ bộ sự liên quan giữa các biến số so với giá chung cư ở thành phố Hồ Chí Minh




Ta thấy rằng diện tích sẽ có sự tương đồng lớn nhất với giá, diện tích căn hộ càng lớn điều đó đồng nghĩa với giá chung cư cũng sẽ nhiều hơn. Tiếp theo đó là phòng ngủ và phòng tắm có sự tương đồng ngang nhau.

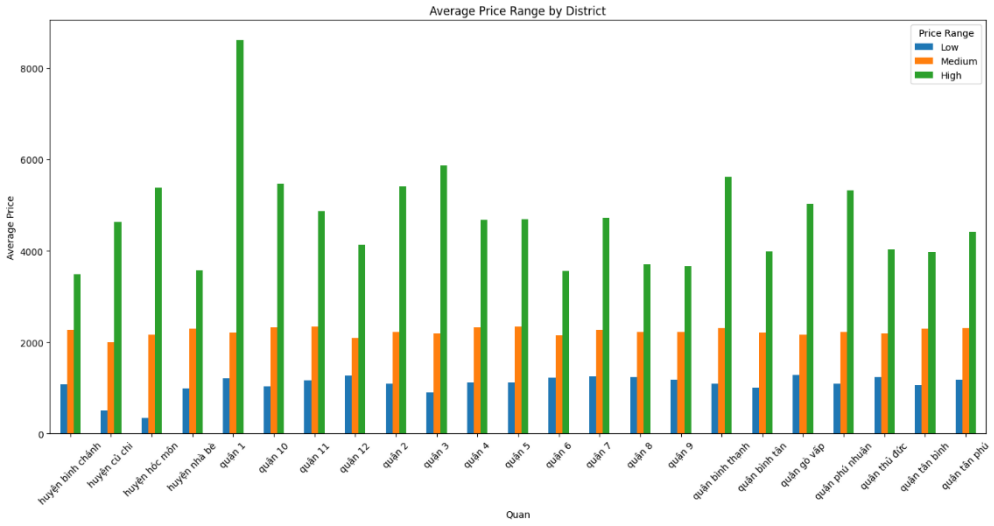



Từ ảnh này cũng cho ta thấy rằng số phòng ngủ và số phòng tắm của một căn chung cư gần như là bằng nhau.

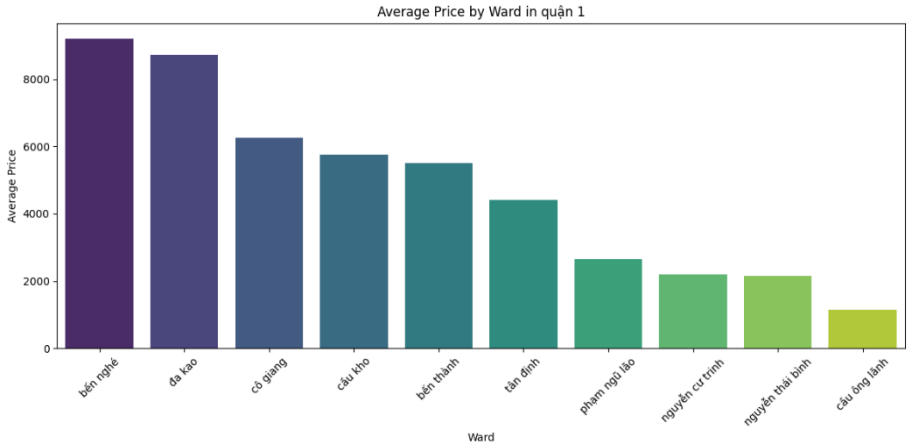



Dưới đây là biểu đồ trung bình giá chung cư tùy theo số phòng tắm và số phòng ngủ càng thể hiện rõ hơn ta thấy rằng càng nhiều phòng tắm và ngủ giá tiền chung cư càng tăng

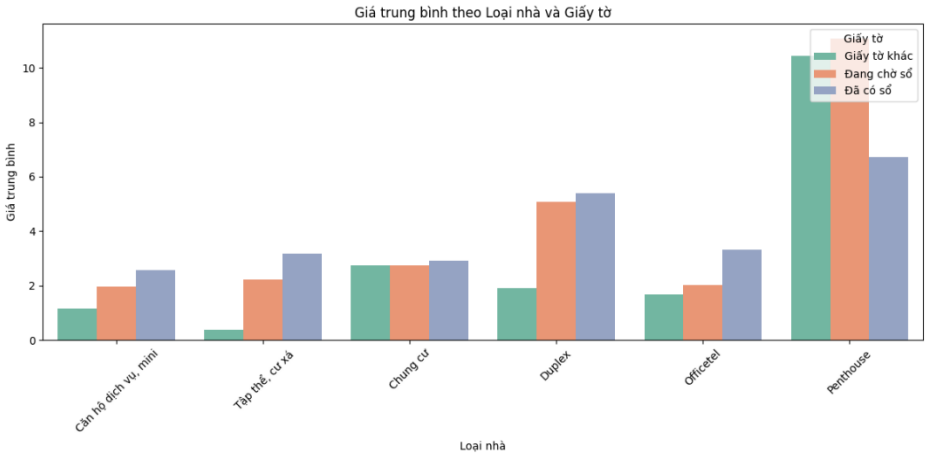
Tương tự chúng em tiến hành sơ bộ với các biến phân loại, nổi bật trong các biến phân loại là các biến về quận



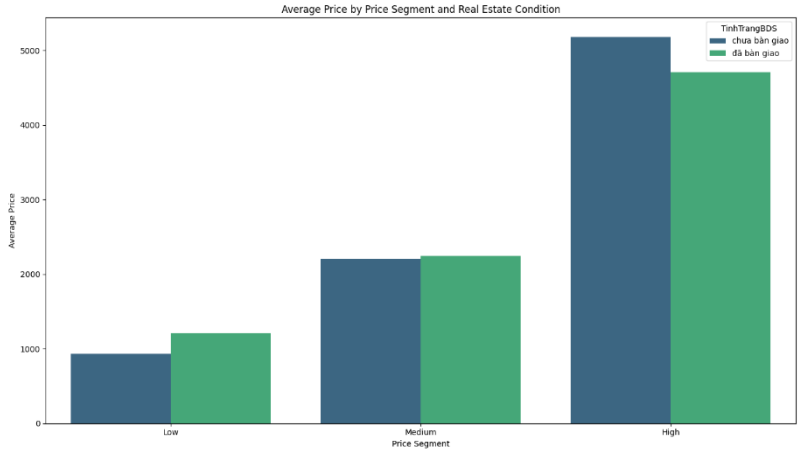

Ta thấy rằng các quận trung tâm và có mật độ dân cư cao như Quận 1, Quận 2 … có giá chung cư ở các phân khúc đều cao hơn hẳn so với các khu vực còn lại.



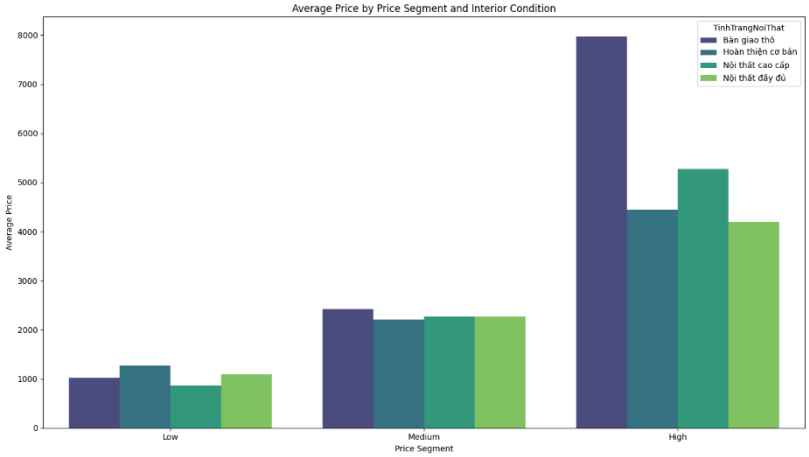

Thêm vào đó không phải trong một quận ở đâu cũng giống nhau, ta thấy rằng trong Quận 1 thì các phường như bến nghé, đa kao sẽ có giá cao nhất và ngược lại cầu ông lãnh sẽ có giá chung cư thấp nhất.



Tiếp theo tới thuộc tính loại nhà và giấy tờ, ta thấy rằng các phân khúc penthouse luôn luôn có giá trị lớn nhất với mọi loại giấy tờ so với các phân khúc còn lại và hầu như đã có sổ luôn có số tiền cao hơn hoàn so với các loại giấy tờ khác trừ loại nhà penthouse.




Ta thấy rằng tình trạng bất sản không ảnh hưởng tới giá thuê nhà quá nhiều, ở phân khúc thấp thì sẽ ưa chuộng tình trạng đã ban giao, nhưng đối với phân khúc cao họ sẽ muốn tự thiết kế vì thế chưa bàn giao giúp họ chủ động hơn trong việc thiết kế vì thế giá tiền sẽ cao hơn

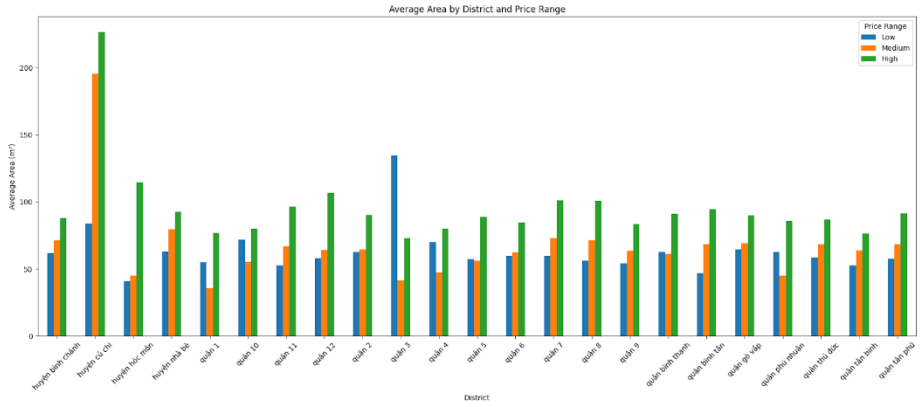



Tương tự với thuộc tính tình trạng nội thất, giá tiền ở phân khúc thấp và trung không ảnh hưởng quá nhiều, nhưng đối với phân khúc cao thì sẽ ưa chuộng bàn giao thô hơn. Có vẻ phân khúc cao ở bàn giao thô sẽ giúp thiết kế nội thất và khung gian sống cho phù hợp nên vì thế giá tiền sẽ cao hơn.

## 5. KẾT QUẢ PHÂN TÍCH

a. Mối quan hệ giữa diện tích và giá chung cư

Kết quả phân tích cho thấy diện tích căn hộ có sự tương đồng rất lớn với giá chung cư. Căn hộ có diện tích lớn hơn thường đi kèm với giá cao, điều này phản ánh xung hướng chung rằng diện tích càng rộng thì diện tích căn hộ càng cao. Biểu đồ trung bình giá chung cư theo diện tích đã chứng minh điều đó. Tuy nhiên, diện tích theo phân khúc của các quận cũng khác nhau.




Những quận xa trung tâm thì đối với diện tích từng phần khúc sẽ càng rộng hơn, nổi bật là huyện củ chi với diện tích căn chung cư rất rộng và trái ngược lại thì quận 1 là quận có diện tích chung cư trung bình thấp nhất ở các phân khúc. Điều này phải ánh tình trạng dân cư tập trung đông đúc ở trung tâm và số tiền phải trả cho một m2 ở nội thành cũng mắt hơn rất nhiều.

b. Sự tương đồng giữa số phòng ngủ, phòng tắm và giá

Khi so sánh phòng ngủ và phòng tắm, ta thấy rằng cả hai biến số này có sự tương đồng ngang nhau trong việc ảnh hưởng tới giá chung cư. Biểu đồ phân tích giá chung cư theo số phòng ngủ và số phòng tắm cho thấy rằng càng nhiều phòng ngủ và phòng tắm, giá căn hộ sẽ càng cao. Đặc biệt, số phòng ngủ và số phòng tắm của các căn hộ thường có mối liên hệ gần như bằng nhau, điều này cho thấy sự cân đối trong thiết kế các căn hộ chung cư

c. Ảnh hưởng của quận và phường đến giá chung cư

Các biến phân loại về quận cho thấy quận trung tâm và có mật độ dân cư cao như Quận 1, Quận 2… đều có giá chung cư cao hơn so với các khu vực khác.

Điều này phản ánh rằng các khu vực trung tâm thường có nhu cầu lớn hơn và giá đất cao hơn, dẫn đến giá căn hộ cao hơn

Tuy nhiên, tùy vào sự quy hoạch và cơ sở vật chất của từng phường trong một quận cũng có sự ảnh hưởng đến giá chung cư. Như ví dụ đã nêu trên, trong quận 1 các phường như Bến Nghé và Đa Kao có giá chung cư cao nhất và ngược lại Cầu Ông Lãnh có giá thấp nhất.

d. Ảnh hưởng của loại nhà và giấy tờ

Như phân tích trên cho thấy tùy theo phân khúc cũng có những giá trị chung cư khác nhau. Penthouse và duplex là 2 loại có giá chung cư cao hơn so với phần còn lại bất kể loại giấy tờ là gì. Hơn nữa, các căn hộ đã có sổ hoặc giấy tờ hợp pháp luôn có giá trị cao hơn so với các căn hộ chưa có sổ. Điều này phản ánh tâm lý của người mua khi ưu tiên các căn hộ có giấy tờ hợp pháp và rõ ràng để đảm bảo quyền sở hữu và sinh sống.

e. Tình trạng bất động sản và giá thuê

Tình trạng bất động sản không ảnh hưởng quá nhiều đến giá thuê nhà, đặc biệt là trong các phân khúc thấp và trung. Tuy nhiên, ở phân khúc cao, khách hàng thường muốn tự thiết kế và trang trí nội thất, vì vậy họ sẽ ưa chuộng các căn hộ chưa bàn giao để có sự chủ động trong thiết kế. Do đó, giá của các căn hộ này có xu hướng cao hơn

f. Ảnh hưởng các tình trạng nội thất đến giá

Tình trạng nội thất cũng không ảnh hưởng quá nhiều đến giá ở các phân khúc thấp và trung. Tuy nhiên, đối với phân khúc cao, các căn hộ bàn giao thô (chưa hoàn thiện nội thất) có xu hướng được ưa chuộng hơn. Điều này cho phép chủ sở hữu tự thiết kế và tạo ra không gian sống phù hợp với nhu cầu cá nhân, vì vậy giá của các căn hộ này cao hơn

Dựa trên các phân tích trên, các yếu tố như diện tích, vị trí quận/phường, loại nhà và giấy tờ tình trạng bất động sản và tình trạng nội thất đều có ảnh hưởng rõ

rệt đến giá chung cư tại thành phố Hồ Chí Minh. Trong đó, diện tích và vị trí địa lý và yếu tố quan trọng nhất. Các căn hộ penthouse và đã có sổ thường có giá cao nhất, trong khi các căn hộ chưa bàn giao hoặc bàn giao thô thường được ưa chuộng trong phân khúc cao cấp.

# 6. XÂY DỰNG MÔ HÌNH MACHINE LEARNING & DEEP LEARNING

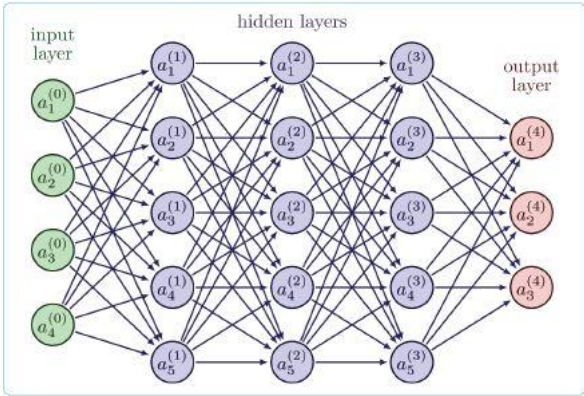
## 6.1. Neural Network

### a. Neural Network là gì?

**Neural Network (Mạng Nơron Nhân Tạo)** là một mô hình học máy dựa trên cấu trúc của não người, với các nút (nơron) được kết nối với nhau thông qua các trọng số. Mạng nơron có thể học các mối quan hệ phức tạp trong dữ liệu và được sử dụng rộng rãi trong các bài toán như nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên, và dự đoán giá trị liên tục (hồi quy).

Cấu  trúc cơ bản của Neural Network :

- Input Layer (lớp đầu vào) : Nhập dữ liệu đầu vào.
- Hidden Layers (Lớp ẩn) : Các lớp trung gian, mỗi lớp chứa nhiều neural. Các neural trong lớp này liên kết với các neural của lớp trước và sau đó thông qua các trọng số.
- Output Layer (Lớp đầu ra) : Cung cấp kết quả dự đoán. Với bài toán dự đoán nhà, lớp đầu ra là một neural với hàm kích hoạt tuyến tính để dự đoán giá trị liên tục.



**b. Tại sao nên chọn Neural Network cho dự đoán giá nhà ?**

- Khả năng học quan hệ phi tuyến tính : Neural Network có khả năng học các mối quan hệ phi tuyến tính giữa các biến đầu vào và đầu ra, điều này rất quan trọng trong việc dự đoán giá nhà, vì giá nhà phụ thuộc vào nhiều yếu tố phức tạp như vị trí, diện tích, số phòng, tiện nghi…
- Khả năng xử lý dữ liệu phức tạp : Mạng neural có thể xử lý cả dữ liệu số và dữ liệu danh mục (categorical) tốt hơn các mô hình đơn giản hơn như hồi quy tuyến tính
- Khả năng mở rộng : Neural Network có thể được mở rộng và tinh chỉnh để cải thiện hiệu suất dự đoán. Bạn có thể thêm các lớp ẩn, thay đổi số lượng mạng neural, hoặc sử dụng các kỹ thuật như Dropout và Batch Normalization để ngăn ngừa overfitting
- Tích hợp với các kĩ thuật khác : Neural Network có thể kết hợp với các kỹ thuật khác như tăng cường học sâu (Deep Learning), giúp tăng cường khả năng dự đoán của mô hình

**c. Mô hình Neural Network**

Mô hình được xây dựng bằng cách sử dụng Sequential từ tensorflow.keras.models cho phép dễ dàng tạo ra một mạng neuron theo từng lớp

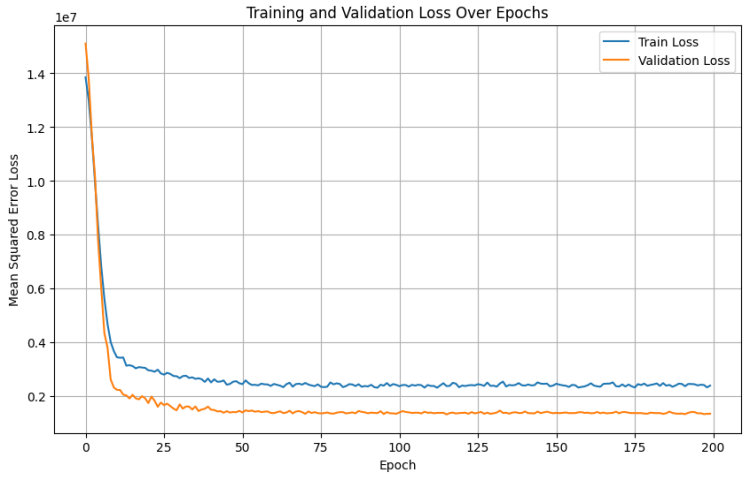
Model: "sequential"

| Layer (type) | Output Shape | Param # |
|---|---|---|
| dense (Dense) | (None, 256) | 46,848 |
| batch_normalization (BatchNormalization) | (None, 256) | 1,024 |
| leaky_re_lu (LeakyReLU) | (None, 256) | 0 |
| dropout (Dropout) | (None, 256) | 0 |
| dense_1 (Dense) | (None, 128) | 32,896 |
| batch_normalization_1 (BatchNormalization) | (None, 128) | 512 |
| leaky_re_lu_1 (LeakyReLU) | (None, 128) | 0 |
| dropout_1 (Dropout) | (None, 128) | 0 |
| dense_2 (Dense) | (None, 64) | 8,256 |
| batch_normalization_2 (BatchNormalization) | (None, 64) | 256 |
| leaky_re_lu_2 (LeakyReLU) | (None, 64) | 0 |
| dropout_2 (Dropout) | (None, 64) | 0 |
| dense_3 (Dense) | (None, 1) | 65 |

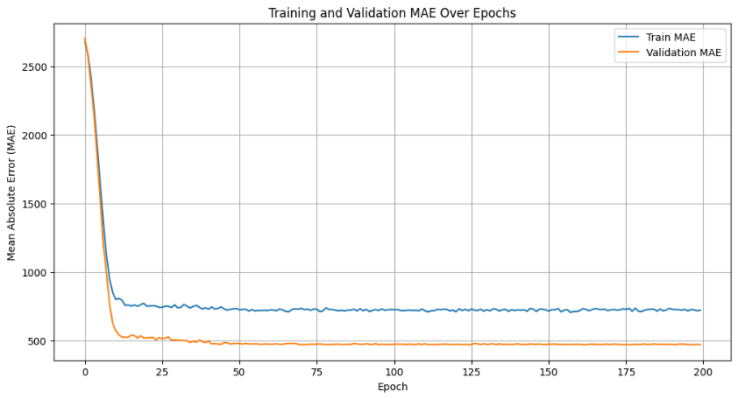
- **Lớp đầu vào :** có 256 neural có nhiệm vụ nhận đầu vào từ dữ liệu đã được biến đổi và truyền qua mạng. Activation = ‘linear’ được sử dụng để giữ nguyên giá trị đầu ra. Điều này là cần thiết cho bài toán hồi quy, nơi cần dự đoán cho một giá trị liên tục
- **Batch Normalization :** Giúp chuẩn hóa đầu ra của lớp hiện tại bằng cách điều chỉnh trung bình và độ lệch chuẩn. Điều này giúp tăng tốc quá trình huấn luyện và giảm hiện tượng overfitting
- **Leaky ReLU :** là một phiên bản cải tiến của hàm ReLU (Rectified Linear Unit), với một độ dốc nhỏ (α) khi đầu vào là giá trị âm, điều này giúp tránh vấn đề "chết nơron" (dead neuron) mà ReLU thường gặp.
- **Dropout :** Là một kỹ thuật regularization để tránh overfitting. Trong quá trình huấn luyện, một tỷ lệ nhất định của các nơron (40% trong trường hợp này) sẽ bị "bỏ qua" (set to zero) để ngừng việc dựa vào các nơron cụ thể.
- **Lớp ẩn :** lớp ẩn thứ 2 và thứ 3 lần lượt giảm số lượng neural từ 265 xuống 128 và 128 xuống 64 để tối ưu bộ tham số và giữa được sự phức tạp của mô hình ở mức hợp lý
- **Lớp đầu ra :** dự đoán giá trị duy nhất, phù hợp cho bài toán hồi quy

**d. Kết quả mô hình thu được**

Biểu đồ này biểu diễn sự thay đổi của Train Loss và Validation Loss theo số lượng Epochs trong quá trình huấn luyện mạng nơron. Nó giúp đánh giá hiệu suất của mô hình và khả năng tổng quát hóa.




Ta thấy rằng Validation Loss giảm đồng bộ với Train Loss cho thấy rằng mô hình đang học tốt và không bị overfitting với tập train. Và mô hình dừng giảm ở khoảng 100 cho thấy mô hình đã được tối ưu.



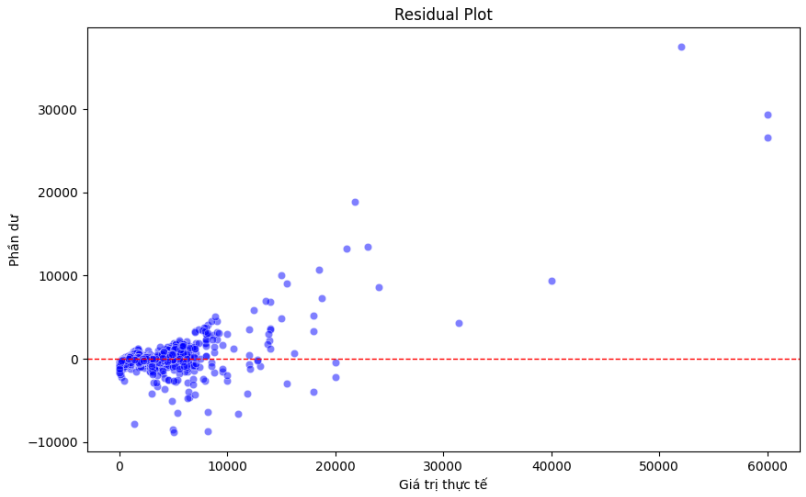

Tương tự đối với Mean Absolute Error (MEA) qua từng Epoch giảm đồng thời chứng tỏ mô hình học tốt, không bị overfitting hay underfitting.

Giá trị trả tối ưu trả về cho thấy :

| | Test Dataset |
|---|---|
| MEA | 477.46 |
| RMSE | 1325.23 |

Sai số trung bình trong khoảng 460 triệu so với giá trị chung cư của toàn tập dữ liệu là ~ 3 tỷ. Cho thấy mô ở mức trung bình tuy nhiên cần cải thiện nhiều hơn về thuật toán để có được kết quả tốt hơn.

Để trực quan hơn dưới đây là biểu đồ Residual Plot:




Ta thấy rằng các giá trị bám khá xác với  đường giá trị thực thực tế, nhưng từ khi giá chung cư từ 10 tỷ trở lên giá trị bắt đầu có sự sai số rộng hơn và có xu hướng đi lên.

## 6.2. Linear Regression

### a. Linear Regression là gì?

Linear Regression (Hồi quy tuyến tính) là một phương pháp thống kê để mô hình hóa mối quan hệ giữa một biến phụ thuộc $y$ (biến mục tiêu) và một hoặc nhiều biến độc lập $x$ (biến dự báo). Mục tiêu của hồi quy tuyến tính là tìm ra một đường thẳng tốt nhất (hoặc một siêu phẳng nếu có nhiều biến dự báo) để dự đoán giá trị của $y$ dựa trên $x$.

### b. Tại sao dùng Linear Regression cho bài toán hồi quy?

1 Đơn giản và dễ hiểu:

- Hồi quy tuyến tính là một trong những phương pháp cơ bản nhất. Kết quả có thể được diễn giải rõ ràng thông qua các hệ số hồi quy β\betaβ, cho biết ảnh hưởng của từng biến đầu vào đến đầu ra.

2 Hiệu quả với dữ liệu tuyến tính:

- Nếu mối quan hệ giữa biến phụ thuộc và biến độc lập là tuyến tính, mô hình này hoạt động rất tốt và cho kết quả chính xác.

3 Nhanh chóng và hiệu quả tính toán:

- Linear Regression có thời gian tính toán nhanh vì sử dụng công thức giải đóng (closed-form solution). Điều này làm cho nó phù hợp với các bài toán lớn hoặc khi cần thử nghiệm nhanh các ý tưởng.

4 Khả năng khái quát hóa tốt với dữ liệu đơn giản:

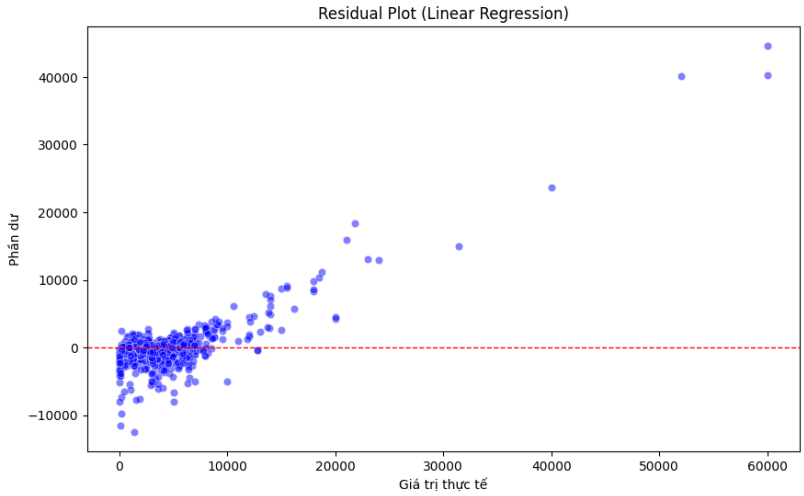
- Với các bài toán hồi quy cơ bản, Linear Regression thường cung cấp kết quả đủ tốt mà không cần đến các mô hình phức tạp hơn.

5 Dễ triển khai và mở rộng:

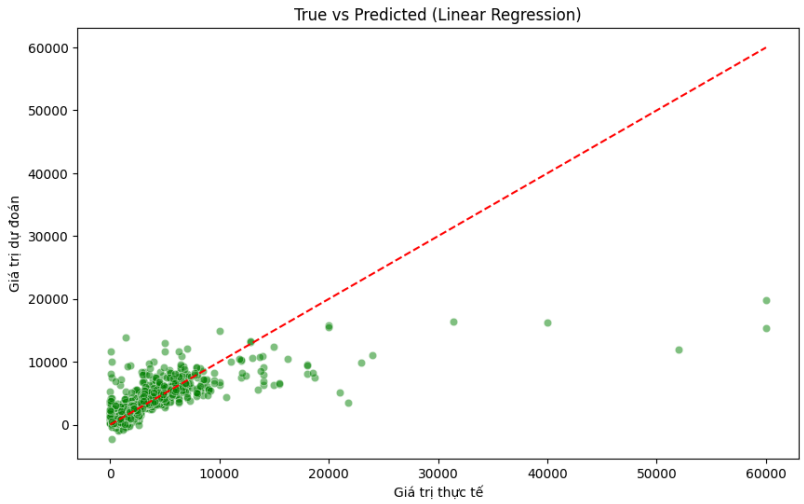
- Linear Regression là cơ sở cho nhiều mô hình khác, như Ridge Regression, Lasso Regression, và Polynomial Regression.

## c. Kết quả đạt được

Để trực quan hơn dưới đây là biểu đồ Residual Plot:




Ta thấy rằng các giá trị bám khá xác với đường giá trị thực thực tế, nhưng từ khi giá chung cư từ 10 tỷ trở lên giá trị bắt đầu có sự sai số rộng hơn và có xu hướng đi lên.




Tiếp theo là biểu đồ tuyến tính so sánh giữa giá trị thực tế và giá trị dự đoán, ta thấy rằng giá trị dự đoán phân bố xung quanh đường tuyến tính. Điều đó chứng tỏ giá dự đoán lệch ít với giá thực tế khi giá nhà dưới 10 tỷ trở xuống và bắt đầu có sai số lớn hơn khi giá nhà trên 20 tỷ.

Bảng kết quả

| | Test Dataset |
|---|---|
| MEA | 774.9 |
| RMSE | 1841.34 |

### 6.3. Gradient Boosting Machines

#### a. Gradient Boosting Machines là gì?

Gradient Boosting Machines (GBM) là một kỹ thuật học máy thuộc nhóm các phương pháp ensemble learning. GBM kết hợp nhiều mô hình đơn giản (thường là cây quyết định) để tạo ra một mô hình mạnh hơn, có khả năng dự đoán chính xác hơn.

GBM hoạt động theo nguyên tắc tăng cường gradient:

1. Mô hình được xây dựng theo cách lặp, mỗi mô hình con (base learner) được huấn luyện để sửa lỗi của mô hình trước đó.
2. Sai số được đo lường thông qua một hàm mất mát (loss function), và mô hình mới được thêm vào để tối thiểu hóa hàm mất mát đó.
3. Quá trình này tiếp tục cho đến khi đạt được số lượng mô hình nhất định hoặc khi sai số hội tụ.

#### b. Tại sao dùng Gradient Boosting Machines trong bài toán hồi quy?

GBM là một lựa chọn mạnh mẽ trong bài toán hồi quy vì:

1. Hiệu suất cao:
    - GBM thường hoạt động rất tốt trên các bài toán hồi quy vì nó tối ưu hóa dần dần dựa trên hàm mất mát, giúp giảm thiểu sai số một cách hiệu quả.
2. Khả năng mô hình hóa quan hệ phức tạp:
    - GBM không yêu cầu mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc. Nó có thể nắm bắt các mô hình phi tuyến tính và tương tác giữa các biến.
3. Linh hoạt với nhiều loại dữ liệu:
    - GBM có thể làm việc tốt với cả dữ liệu dạng số, dạng danh mục, và xử lý tốt các dữ liệu mất mát sau khi tiền xử lý.
4. Khả năng tùy chỉnh cao:
    - Có nhiều siêu tham số (hyperparameters) để điều chỉnh, như số lượng cây, chiều sâu của cây, learning rate, v.v., giúp tối ưu hóa hiệu suất trên các bài toán cụ thể.
5. Chống overfitting tốt hơn (với regularization):

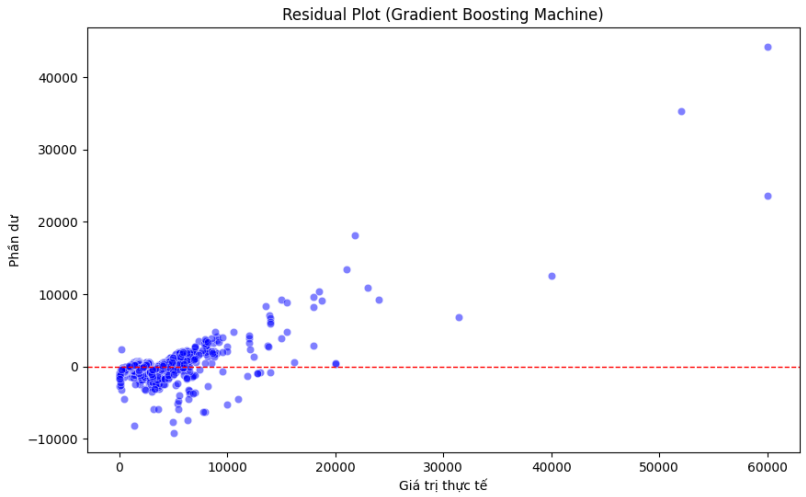
- Các cơ chế như learning rate thấp và số lượng cây giới hạn giúp giảm nguy cơ overfitting.

6. Ứng dụng rộng rãi:
   - GBM đã được chứng minh hiệu quả trong nhiều bài toán thực tế như dự đoán giá, dự báo doanh số, và phân tích rủi ro.

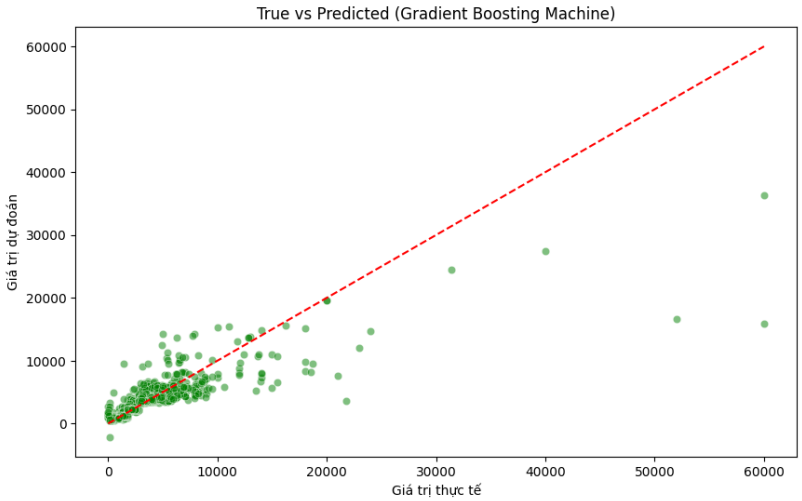
### c. Kết quả đạt được

Để trực quan hơn dưới đây là biểu đồ Residual Plot:

Ta thấy rằng các giá trị bám khá xác với  đường giá trị thực thực tế, nhưng từ khi giá chung cư từ 10 tỷ trở lên giá trị bắt đầu có sự sai số rộng hơn và có xu hướng đi lên.




Tiếp theo là biểu đồ tuyến tính so sánh giữa giá trị thực tế và giá trị dự đoán, ta thấy rằng giá trị dự đoán phân bố xung quanh đường tuyến tính. Điều đó chứng tỏ giá dự đoán lệch ít với giá thực tế khi giá nhà dưới 10 tỷ trở xuống và bắt đầu có sai số lớn hơn khi giá nhà trên 20 tỷ.




Bảng kết quả

| | Test Dataset |
|---|---|
| MEA | 579.21 |
| RMSE | 1493.15 |

### 6.4. So sánh và đánh giá

Dưới đây là bảng kết quả đánh giá trên cùng một tập Test của 3 mô hình Neural Network, Linear Regression và Gradient Boosting Machines (GBM) dựa trên 2 chỉ số MAE (Mean Absolute Error) và RMSE (Root Mean Square Error).

| | Neural Network | Linear Regression | Gradient Boosting Machines |
|---|---|---|---|
| MEA | 477.46 | 774.9 | 579.21 |
| RMSE | 1325.23 | 1841.34 | 1493.15 |

⇨ Mô hình Neural Network cho kết quả tốt nhất (MEA thấp nhất (477.46) và RMSE (1325.23) ở mức khá). GBM cân bằng giữa hiệu suất và độ phức tạp, GBM đạt MEA (579.21) và RMSE (1493.15). Linear Regression đạt MEA (774.9) và RMSE (1841.34) cao nhất trong ba mô hình, cho thấy hiệu quả kém nhất trong 3 mô hình.

## 7. KẾT LUẬN

Từ bộ dữ liệu về thông tin các căn chung cư được bán trên địa bàn thành phố Hồ Chí Minh năm 2020 được thu thập trên trang Chợ tốt[2]. Để giải đáp câu hỏi "Yếu tố nào ảnh hưởng nhiều nhất đến giá chung cư trên địa bàn thành phố Hồ Chí Minh?" từ bộ dữ liệu đã được chuẩn hoá nhóm đã tiến hành phân tích các mối quan hệ của các biến độc lập với giá chung cư, kiểm chứng các giả định. Từ đó nhóm rút ra các Insight quan trọng như: Diện tích căn hộ có sự tương đồng rất lớn đến giá chung cư; Các quận trung tâm và có mật độ dân cư cao như Quận 1, Quận 2,.. đều có giá chung cư cao hơn với các khu vực khác; Tuỳ theo phân khúc giá chung cư sẽ khác nhau nhóm nhận thấy Penthouse và duplex là 2 loại có giá chung cư cao hơn so với phần còn lại;… Thông qua những hiểu biết đó nhóm sử dụng các thuộc tính quan trọng xây dựng 3 mô hình dự đoán chính là Neural Network, Linear Regression và mô hình Ensemble sử dụng kỹ thuật Gradient Boosting Machines. Sau quá trình tinh chỉnh, đánh giá, mô hình có kết quả dự đoán tốt nhất là Neural Network với chỉ số trên tập Test: MEA là 477.46 và RMSE là 1325.23. Vì bộ dữ liệu giá nhà thu thập từ năm 2020, do giá căn hộ chung cư có thể đã biến động theo sự thay đổi của thị trường đến thời điểm hiện tại, mô hình dự đoán có thể gặp sai số khi áp dụng cho các mẫu dữ liệu giá nhà hiện nay.

# TÀI LIỆU THAM KHẢO

[1] Github. Link: https://github.com/HungTrinhIT/FinalProject-Datascience.git. (25/10/2024)

[2] Chợ tốt. Link: https://www.chotot.com/ (25/10/2024).

[3] Scikit-learn.org. Link: https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestClassifier.html (06/12/2024)

[4] IBM. Link: https://www.ibm.com/topics/neural-networks (06/12/2024)

[5] Scikit-learn.org. Link: https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LinearRegression.html (06/12/2024)

[6] Scikit-learn.org. Link: https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (06/12/2024)

# PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

| STT | Thành viên | Nhiệm vụ |
|---|---|---|
| 1 | Nguyễn Tuệ Minh | Trách nghiệm: Thành viên<br>Công việc:<br>- Tìm hiểu bộ dữ liệu<br>- Phân tích thăm dò dữ liệu<br>- Trực quan hoá dữ liệu<br>- Viết báo cáo<br>- Chỉnh báo cáo |
| 2 | Nguyễn Phúc Khang | Trách nhiệm: Thành viên<br>Công việc:<br>- Tìm hiểu bộ dữ liệu<br>- Xử lý dữ liệu thô<br>- Phân tích thăm dò dữ liệu<br>- Trực quan hoá dữ liệu<br>- Huấn luyện mô hình<br>- Viết báo cáo<br>- Làm slide |
| 3 | Đỗ Nguyễn Anh Khoa | Trách nhiệm: Nhóm trưởng<br>Công việc:<br>- Tìm hiểu bộ dữ liệu<br>- Phân tích thăm dò dữ liệu<br>- Trực quan hoá dữ liệu<br>- Huấn luyện mô hình<br>- Viết báo cáo<br>- Chỉnh source code |