ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

BÁO CÁO ĐỒ ÁN MÔN HỌC

ĐỀ TÀI: Final Projects - House Price Prediction - 22KDL

***

Giảng viên hướng dẫn	**: Ngô Minh Mẫn**	
	Lê Hoàng Đức	
Lớp	**: 22KDL**	
Danh sách sinh viên thực hiện	**: Lê Quốc An**	**- 22280001**
	Nguyễn Công Tiến Dũng	**- 22280014**
	Lư Xuân Dương	**- 22280015**
	Nguyễn Đức Hiệp	**- 22280022**

Thành phố Hồ Chí Minh, 21 tháng 6 năm 2024

MỤC LỤC

GIỚI THIỆU

Đặt vấn đề

Việc mua bán nhà chưa bao giờ là hết phổ biến nhưng để có thể đưa ra được nhận định giá cho người bán và người mua, hay hỗ trợ trong các tư vấn tài chính như các khoản vay giúp giảm thiểu rủi ro trong đầu tư bất động sản thì mỗi cá nhân rất khó để phân tích và xử lí một nguồn dữ liệu lớn và đưa ra quyết định hợp lí.

Để giải quyết được vấn đề đó thì việc xây dựng một mô hình dự đoán giá nhà có thể giúp cho định giá chính xác hơn từ việc thu thập các dữ liệu bất động sản thu thấp được trên website về rao bán bất động sản.

Mục tiêu đồ án

Thu thập và xử lý dữ liệu: Mục tiêu đầu tiên của đồ án là thu thập dữ liệu liên quan đến bất động sản (cụ thể là giá nhà) ở trên khắp Việt Nam từ website : batdongsan sau đó chúng em xử lý dữ liệu để phục vụ cho việc phân tích và trực quan hóa.

Yêu cầu đồ án

- Thu thập dữ liệu:

+ Cào và thu thập các mục dữ liệu từ website https://batdongsan.vn/ban-nha/ để tạo thành bộ dữ liệu toàn diện.

+ Khám phá trang web một cách kỹ lưỡng để thu thập thông tin liên quan trên từng mục dữ liệu để dự đoán giá nhà.

+ Xác định kích thước của tập dữ liệu cần thu thập.

- Xử lý trước dữ liệu và kỹ thuật tính năng:

+ Chọn các tính năng dữ liệu thích hợp để đào tạo các mô hình.

+ Thực hiện các bước tiền xử lý dữ liệu cần thiết như xử lý các giá trị bị thiếu, mã hóa các biến phân loại và chia tỷ lệ các tính năng số.

+ Tiến hành các kỹ thuật kỹ thuật tính năng để nâng cao khả năng dự đoán của các tính năng đã chọn.

- Phương pháp:

+ Trực quan hóa bản đồ nhiệt dựa trên địa chỉ của ngôi nhà được liệt kê bằng cách sử dụng ước tính mật độ hạt nhân tại một thời điểm nhất định.

+ Thiết kế một phương pháp để giải quyết vấn đề dự đoán giá nhà.

+ Xem xét các phương pháp và kỹ thuật khác nhau để phân tích và lập mô hình dữ liệu.

+ Chứng minh các phương pháp đã chọn và giải thích tính hiệu quả của chúng.

+ Các mô hình được đề xuất bao gồm Linear/Ridge/Lasso Regression, Decision Tree/Random Forest, and Gradient Boosting và các mô hình khác.

+ Khuyến khích học sinh khám phá các mô hình nâng cao nếu khả thi.

+ Nếu các mô hình học sâu được sử dụng, hãy ưu tiên và đề xuất framework PyTorch.

- Chọn số liệu có liên quan để đánh giá mô hình.

+ Nếu sử dụng nhiều mô hình, học sinh nên thực hiện các tiêu chuẩn mô hình và xác định mô hình tốt nhất cho vấn đề.

Phân công công việc và đánh giá

Bảng 1 – Bảng phân công công việc

STT	**Nội dung công việc**	**Người phụ trách**
1	Crawl data	Lê Quốc An, Lư Xuân Dương, Nguyễn Đức Hiệp
2	Xử lý dữ liệu	Tất cả các thành viên
3	Tìm hiểu và xây dựng các mô hình máy học phù hợp	Nguyễn Công Tiến Dũng
4	Tổng hợp và thu thập lại kết quả của nhóm	Nguyễn Công Tiến Dũng
5	Viết báo cáo đồ án	Tất cả các thành viên

Bảng 2 - Bảng đánh giá thành viên

	An	Dũng	Dương	Hiệp
An	X	9.8	9.6	9.7
Dũng	9.7	X	9.8	9.6
Dương	9.7	9.5	X	9.8
Hiệp	9.6	9.8	9.7	X
Tổng kết	9.7	9.7	9.7	9.7

*Tất cả các thành viên đều chung sức, trao đổi làm chung đồ án, còn người phụ trách là người phụ trách chính về phần đó cho nên mức độ đánh giá các thành viên nhóm đều sẽ như nhau cho cả 4 bạn

Chương 1: Xác định vấn đề

1.1 - Vấn đề chính:

- **Dự đoán giá nhà:** Mục tiêu chính của đồ án là xây dựng một mô hình máy học có khả năng dự đoán giá nhà dựa trên các thông tin thu thập được từ trang web [invalid URL removed].vn.

1.2- Các vấn đề cụ thể:

1.2.1 - Thu thập dữ liệu:

- **Cào dữ liệu:** Đồ án yêu cầu sinh viên thu thập dữ liệu từ website https://batdongsan.vn/ban-nha/. Đây là một thử thách vì cần phải hiểu rõ cấu trúc website, xử lý các vấn đề liên quan đến việc thay đổi cấu trúc trang, và đảm bảo tính ổn định của quá trình thu thập dữ liệu.
- **Xác định đặc trưng:** Sinh viên cần phải lựa chọn các đặc trưng phù hợp từ dữ liệu thu thập được để đưa vào mô hình dự đoán. Việc này đòi hỏi sự hiểu biết về thị trường bất động sản và các yếu tố ảnh hưởng đến giá nhà.
- **Kích thước bộ dữ liệu:** Cần phải xác định số lượng dữ liệu cần thu thập để đảm bảo đủ thông tin cho việc huấn luyện mô hình và đánh giá hiệu quả.

1.2.2 - Tiền xử lý và kỹ thuật đặc trưng:

- **Xử lý dữ liệu bị thiếu:** Dữ liệu thu thập được có thể chứa các giá trị bị thiếu, cần phải có phương pháp xử lý phù hợp để đảm bảo tính toàn vẹn của dữ liệu.
- **Mã hóa biến phân loại:** Các biến phân loại (như loại nhà, khu vực,...) cần được chuyển đổi thành dạng số để có thể đưa vào mô hình.
- **Chuẩn hóa đặc trưng:** Các đặc trưng số có thể có phạm vi giá trị khác nhau, cần phải được chuẩn hóa để đảm bảo tính công bằng trong quá trình huấn luyện mô hình.
- **Kỹ thuật đặc trưng:** Sinh viên có thể áp dụng các kỹ thuật tạo đặc trưng mới để tăng cường khả năng dự đoán của mô hình.

1.2.3 - Phương pháp luận:

- **Trực quan hóa dữ liệu:** Đồ án yêu cầu sinh viên trực quan hóa dữ liệu bằng heatmap dựa trên địa chỉ nhà, sử dụng kỹ thuật kernel density estimation.
- **Lựa chọn mô hình:** Cần lựa chọn mô hình phù hợp với bài toán dự đoán giá nhà. Có thể dùng các mô hình được xây dựng sẵn hoặc tự xây dựng model bằng Frame work Pytorch
- **Đánh giá mô hình:** Cần phải chọn các metric đánh giá phù hợp để so sánh hiệu suất của các mô hình khác nhau và lựa chọn mô hình tốt nhất.

1.2.3 - Triển khai và báo cáo:

- **Sử dụng Jupyter Notebook:** Sử dụng Jupyter Notebook để triển khai mô hình và viết báo cáo.

Chương 2: QUÁ TRÌNH TRIỂN KHAI ĐỒ ÁN

2.1 - Quy trình thu thập và xử lý dữ liệu

2.1.1 – Chọn website để thu thập

Website lựa chọn là https://batdongsan.vn/ban-nha/ .

2.1.2 – Chọn hình thức thu thập dữ liệu

Sử dụng thư viên selenium của python để crawl dữ liệu từ website. Vì Selenium là một công cụ mạnh mẽ cho phép chúng ta tự động hóa các tác vụ trên trình duyệt web, nó cho phép chúng ta mô phỏng hành động của người dùng như nhấp chuột, nhập liệu, và điều hướng qua các trang web. Vì thế selenium rất phù hợp để crawl dữ liệu

2.2 - Quy trình xử lý dữ liệu

Xử lý dữ liệu bắt đầu với việc thu thập dữ liệu. Dữ liệu ban đầu có cấu trúc HTML nhưng để thống kê và trực qua hóa thì chúng ta nên đưa về dạng csv hoặc excel.

Chi tiết các bước như sau:

Bước 1 – Thu thập dữ liệu: Trong dự án này, chúng em đã sử dụng Selenium để thu thập dữ liệu từ một trang web bất động sản. Mục tiêu là lấy thông tin về các căn nhà được đăng bán bao gồm tên, giá, diện tích, địa chỉ, và các thông tin khác. Chúng em cài đặt môi trường Selenium tiếp theo cài đặt cấu hình và khởi tạo WebDriver tiếp đến lấy thông tin từ trang chủ bằng việc tìm kiếm các phần tử bằng CSS Selectors và Xpath sau đó lấy thông tin từ các trang con khác (là từng ngôi nhà) lấy các thông tin như : diện tích, số phòng ngủ, số phòng wc, hướng nhà, hướng ban công, địa chỉ của ngôi nhà là tỉnh và huyện sau đó tất cả các thông tin của từng ngôi nhà sẽ được lưu trữ vào 1 file csv (xem chi tiết trong file cào dữ liệu của nhóm).
Bước 2 – Xử lý dữ liệu: Chuyển các dữ liệu của các cột về đúng cột của mình tại vì có các cột có các dữ liệu không phải cột của mình. Sau đó loại bỏ các chữ cái trong các cột diện tích, giá, phòng ngủ, phòng wc bởi vì các cột này chúng ta chỉ nên giữ lại dữ liệu dạng số để có thể áp dụng các model vào còn các dữ liệu bị trống khi cào thì ta để thành NaN để chúng ta phân tích sau đó.
Bước 3 – Phân tích dữ liệu: Sau khi xử lý dữ liệu trên tiếp theo ta đổi tên các cột và chỉnh sửa thứ tự của các cột để dễ dàng sử dụng theo sự thống nhất của cả nhóm. Đổi các type của từng cột cho phù hợp với dữ liệu cột đó. Ví dụ cột Giá thì ta chuyển về type float và chuyển nó về giá trị của tỷ, cột Thời gian đăng chuyển thành date, các các cột phòng ngủ, phòng wc chuyển thành float,.....

Bước 4 – Trình bày dữ liệu: Sau khi đã phân tích và xử lý thì sau đó sẽ tổng kết lại dữ liệu xem còn thiếu sót gì không, sau đó đưa dữ liệu vào 1 file csv mới (dataFinal). Chi tiết về file csv mới ở bên dưới

Chương 3: KẾT QUẢ VÀ THẢO LUẬN

3.1 – Dữ liệu sau khi thu thập và xử lý (tại đây)

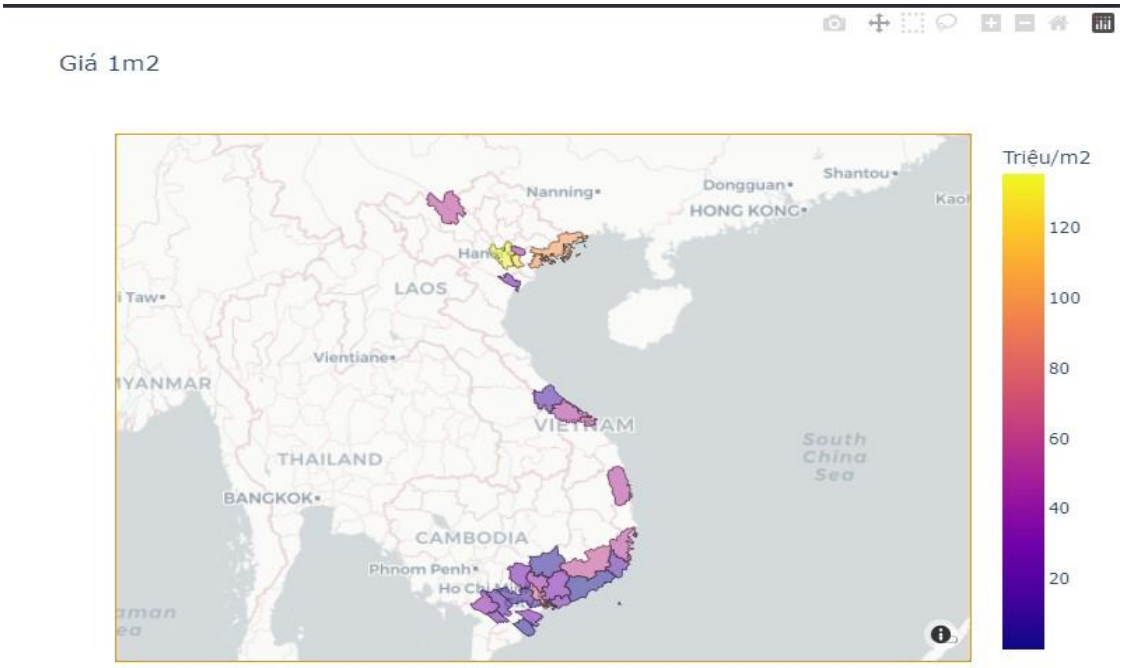
	A	B	C	D	E	F	G	H	I	J	K	L
1	Tiêu đề	Diện tích(m2)	Số phòng ngủ	Số phòng WC	Thời gian đăng	Tỉnh/Thành	Quận/Huyện	Hướng nhà	Hướng ban công	Loại nhà	Giá	Link
2	Chưa tới 30tr/m2 - H	150	2	1	10/12/2023	TP Hồ Chí Minh	Nhà Bè			Bán Nhà riêng	3899000000	https://batdongsan.vn
3	Bán nhà HXH Âu Cơ	51			10/12/2023	TP Hồ Chí Minh	Tân Bình			Bán Nhà riêng	5.5	https://batdongsan.vn
4	SÁT MẶT TIỀN PHA	45	2	2	10/12/2023	TP Hồ Chí Minh	Phú Nhuận			Bán Nhà	4.6	https://batdongsan.vn
5	CHỦ GẤP BÁN TRƯ	41			10/12/2023	TP Hồ Chí Minh	Quận 5			Bán Nhà riêng	7.35	https://batdongsan.vn
6	LŨY BÁN BÍCH,TÂN	96	2	1	07/12/2023	TP Hồ Chí Minh	Tân Phú			Bán Nhà riêng		https://batdongsan.vn
7	Bán nhà trong ngõ L	40	4	3	08/12/2023	Hà Nội	Hà Đông			Bán Nhà riêng	3.95	https://batdongsan.vn
8	Bán nhà trong ngõ L	40	4	3	08/12/2023	Hà Nội	Hà Đông			Bán Nhà riêng	3.95	https://batdongsan.vn
9	BÁN NHÀ PHỐ HOÀI	59			08/12/2023	Hà Nội	Thanh Xuân			Bán Nhà riêng	7.95	https://batdongsan.vn
10	BÁN NHÀ MỚI ĐẸP 7 X 13, CHỈ 3.5 TỶ, H		3	3	08/12/2023	TP Hồ Chí Minh	Thủ Đức			Bán Nhà mặt phố		https://batdongsan.vn
11	SIÊU PHẨM Chào bá	425	6	6	08/12/2023	Hải Phòng	Ngô Quyền	Tây-Nam	Tây-Nam	Bán Nhà mặt phố	17	https://batdongsan.vn
12	Bán Nhà Mặt tiền đu	180			08/12/2023	TP Hồ Chí Minh	Thủ Đức			Bán Nhà	19	https://batdongsan.vn
13	Bán nhà Âu Cơ-Tây	42	4	3	08/12/2023	Hà Nội	Tây Hồ			Bán Nhà riêng	5.29	https://batdongsan.vn
14	Bán nhà phố Đào Tấ	39	4	5	08/12/2023	Hà Nội	Ba Đình			Bán Nhà riêng	8.6	https://batdongsan.vn
15	Trương Đức Toàn	80	8	8	08/12/2023	TP Hồ Chí Minh	Quận 11	Nam	Nam	Bán Nhà mặt phố	21.8	https://batdongsan.vn

Hình 3.1: Dữ liệu sau khi đưa xử lí

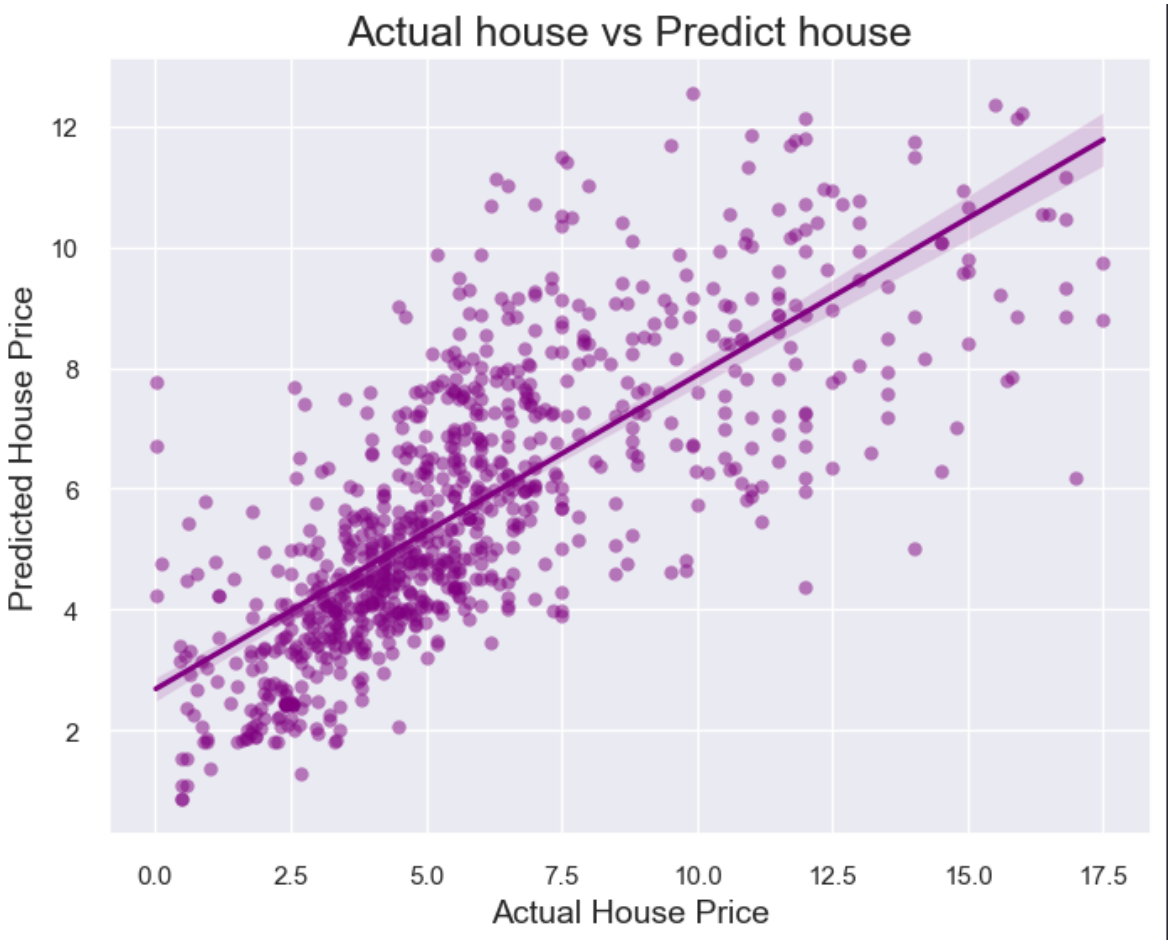
3.2 – Các kết quả thu được từ việc phân tích và sử dụng mô hình máy học

Giá nhà (Giá (triệu) / m^2) dựa theo trung bình các tỉnh thành sau khi chúng em trực quan

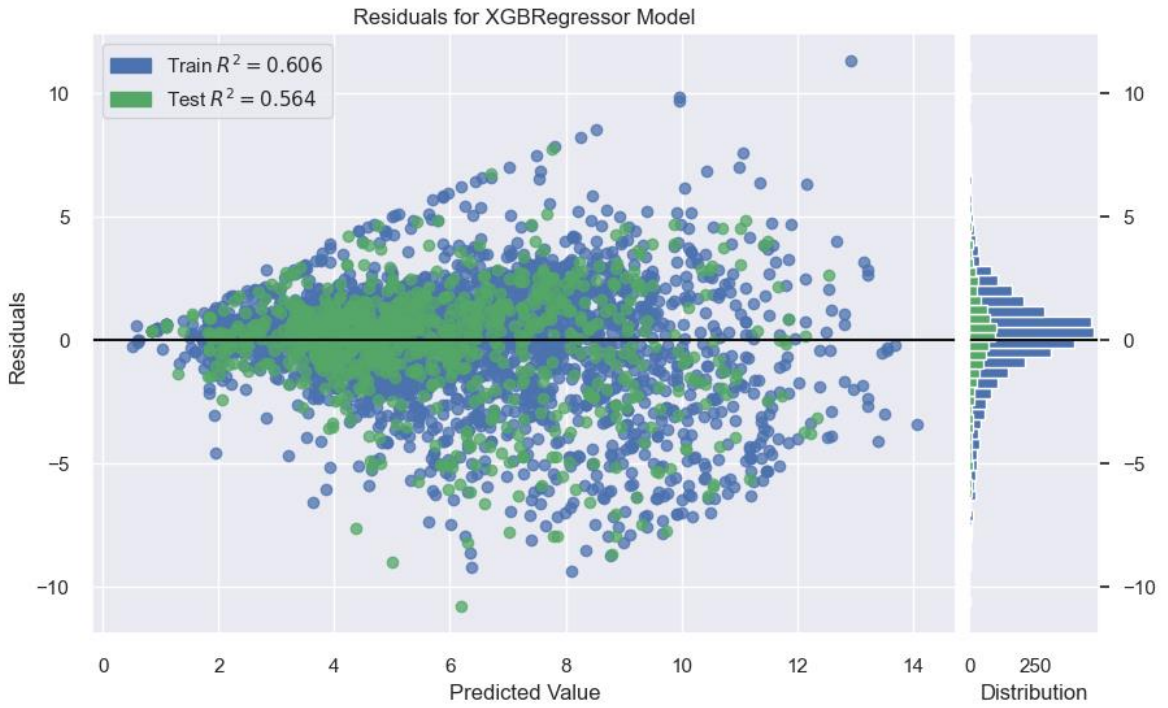
5



Hình 3.2: Bản đồ về giá đất/m2 của các Tỉnh thành trên cả nước có dữ liệu



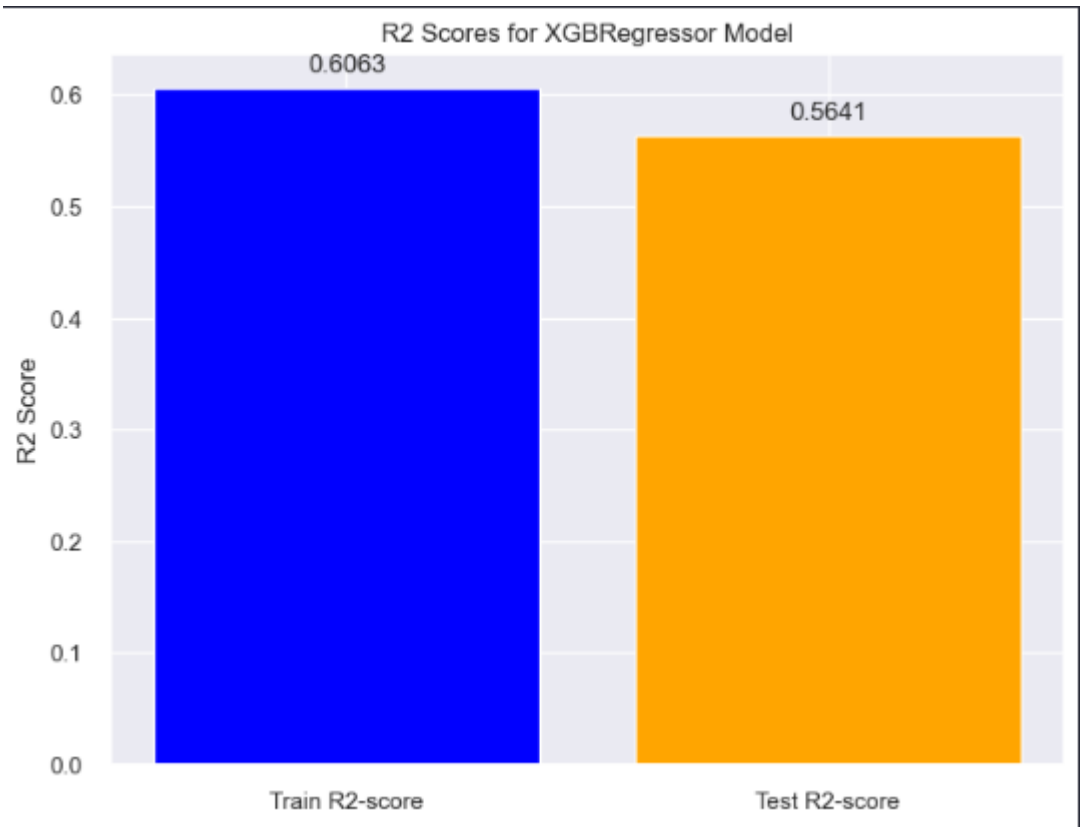
6



Hình 3.3 và 3.4: Kết quả từ việc áp dụng Mô hình hồi quy XGBoost

Từ kết quả thu được thì có thể thấy rằng bộ dữ liệu giá nhà khi crawl trên website bất động sản thì có dữ liệu giá phân bố tuyến tính. Tuy nhiên dữ liệu còn phân tán khá nhiều nên cho ra hiệu xuất của mô hình không được cao

7



Hình 3.5: Hiệu suất của mô hình khi cho ra trên tập train và test được đánh giá bằng R2-score

Có thể thấy đươc rằng hiệu suất mô hình đạt được là tương đối cao so với một bộ dữ liệu có bộ phân tán lớn như thế này.

3.4 - Thảo luận

3.4.1 – Thuận lợi

Sự hướng dẫn kỹ lưỡng của giảng viên trong quá trình thực hiện đồ án là một lợi thế lớn. Những vấn đề khó khăn trong quá trình thực hiện được giải đáp cụ thể và nhanh chóng bởi giảng viên hướng dẫn.

Ngoài ra, việc phân công công việc rõ ràng và hợp lý giữa các thành viên trong nhóm giúp đảm bảo tiến độ và chất lượng công việc. Mỗi thành viên đều có trách nhiệm và nhiệm vụ riêng, góp phần vào sự thành công chung của dự án.

Nhóm có sự hỗ trợ từ các tài liệu và nguồn học liệu phong phú, bao gồm bài báo khoa học, và các tài liệu trực tuyến. Điều này giúp nhóm nắm bắt và áp dụng các kiến thức cần thiết một cách nhanh chóng và hiệu quả.

Sự hợp tác và phối hợp tốt giữa các thành viên trong nhóm cũng là một yếu tố quan trọng. Mọi người đều có tinh thần làm việc nhóm cao, sẵn sàng hỗ trợ nhau khi gặp khó khăn và chia sẻ kiến thức, kinh nghiệm.

Công cụ và công nghệ hiện đại như Google Colab, Jupyter Notebook, và các thư viện Python mạnh mẽ như Pandas, Selenium, và Matplotlib giúp cho quá trình phân tích và trực quan hóa dữ liệu trở nên dễ dàng và hiệu quả hơn. Các công cụ này cung cấp môi trường làm việc linh hoạt và mạnh mẽ, giúp nhóm tiết kiệm thời gian và nâng cao chất lượng sản phẩm.

Nhóm cũng tận dụng được các kho dữ liệu mở và các AI miễn phí để thu thập dữ liệu cần thiết cho dự án. Việc này giúp giảm bớt thời gian và công sức trong việc thu thập dữ liệu từ các nguồn khác nhau.

Sự phản hồi liên tục từ giảng viên và các thành viên khác trong nhóm giúp nhóm điều chỉnh và cải thiện sản phẩm một cách nhanh chóng. Việc này đảm bảo rằng dự án luôn đi đúng hướng và đạt được các mục tiêu đặt ra.

Cuối cùng, tinh thần học hỏi và cầu tiến của các thành viên trong nhóm là một yếu tố quan trọng. Mọi người đều sẵn sàng học hỏi các kỹ năng mới và áp dụng chúng vào dự án, từ đó nâng cao chất lượng và hiệu quả công việc.

3.4.2 – Khó khăn

Việc lựa chọn thông tin để khai thác trên website gặp nhiều vấn đề vì trên trang web có rất nhiều thông tin. Chúng em phải lọc kỹ những thông tin cần thiết để cào về, điều này tốn rất nhiều thời gian. Ngoài ra, việc cào một khối dữ liệu lớn về máy cũng đòi hỏi thời gian và công sức đáng kể.

Sau khi cào dữ liệu xong, dữ liệu thường rất thô và cần xử lý nhiều thứ, từ định dạng của từng cột dữ liệu đến các thông tin trong các cột đó. Ví dụ, cột Giá cần được xử lý để chỉ để lại các con số tượng trưng cho giá tiền của ngôi nhà. Các cột bị trộn lẫn dữ liệu vào nhau cũng phải được xử lý kỹ càng để đưa về đúng với cột của chúng.

Khi trực quan hóa dữ liệu bằng heatmap, chúng em cần phải tìm file geojson phù hợp, có chứa các tọa độ của các tỉnh để có thể trực quan dữ liệu mật độ giá nhà/m^2 của từng tỉnh. Việc tìm kiếm file geojson cũng yêu cầu lọc kỹ từng chi tiết.

Do sự nhầm lẫn về mục tiêu trong đồ án, chúng em đã đi sai hướng về heatmap mà thầy yêu cầu nhưng đã phát hiện và khắc phục vào những ngày cuối làm đồ án.

Sự bất đồng quan điểm về các ý kiến của các thành viên trong nhóm cũng là một khó khăn. Triển khai cào data trên nhiều nền tảng khác nhau như Jupyter Notebook và Google Colab gây ra nhiều mâu thuẫn về output của các feature.

Một số trang web có thể chặn truy cập từ các địa chỉ IP của Google Colab do lo ngại về bảo mật hoặc nhận diện hoạt động crawl dữ liệu. Việc này gây ra hạn chế về quyền truy cập và bảo mật.

Dù Google Colab cung cấp tài nguyên mạnh mẽ, nhưng nó vẫn có giới hạn về thời gian sử dụng và tài nguyên (RAM, CPU). Các phiên làm việc có thể bị ngắt quãng sau một thời gian nhất định.

Việc cài đặt các công cụ cần thiết trên Google Colab tuy dễ dàng, nhưng chúng em cần phải đảm bảo rằng các phiên bản của ChromeDriver và trình duyệt Chrome đều tương thích với nhau và với Selenium.

Google Colab có thể giới hạn số lượng kết nối mạng hoặc băng thông, ảnh hưởng đến tốc độ và hiệu quả của việc crawl dữ liệu. Ngoài ra, Colab có thể ngắt phiên làm việc của chúng em sau một thời gian không hoạt động hoặc khi vượt quá giới hạn sử dụng tài nguyên, gây gián đoạn quá trình crawl dữ liệu liên tục.

Việc debug các vấn đề liên quan đến Selenium trên Google Colab cũng khó khăn hơn so với việc chạy trên máy tính cá nhân, vì chúng em không thể tương tác trực tiếp với trình duyệt. Điều này làm cho việc kiểm tra và sửa lỗi trở nên phức tạp hơn.

Một khó khăn khác là việc tối ưu hóa hiệu suất của các script crawl. Việc viết code không tối ưu có thể dẫn đến thời gian chạy dài hơn và sử dụng tài nguyên không hiệu quả. Điều này đặc biệt quan trọng khi phải xử lý một lượng lớn dữ liệu trong thời gian ngắn.

Việc tạo ra các bộ dữ liệu sạch và đáng tin cậy từ dữ liệu thô cũng đòi hỏi kỹ năng xử lý dữ liệu nâng cao. Chúng em phải sử dụng các kỹ thuật làm sạch dữ liệu, loại bỏ các giá trị ngoại lai, và chuẩn hóa dữ liệu để đảm bảo tính nhất quán và chính xác.

Cuối cùng, việc lập kế hoạch và quản lý thời gian hiệu quả là một thách thức lớn. Chúng em phải cân đối giữa việc học tập, làm việc nhóm và thời gian cá nhân để đảm bảo

Chương 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

4.1 - Kết luận

Đồ án dự đoán giá nhà đã mang lại những bài học quý giá về cả kỹ thuật lẫn nghiệp vụ.

Về mặt kỹ thuật, việc lựa chọn mô hình phù hợp (như XGBoost hoặc hồi quy tuyến tính) và tinh chỉnh siêu tham số đóng vai trò quan trọng trong việc tối ưu hóa hiệu suất dự đoán. Việc xử lý overfitting bằng các kỹ thuật như regularization và cross-validation cũng không thể bỏ qua để đảm bảo mô hình có khả năng tổng quát hóa tốt trên dữ liệu mới. Việc đánh giá mô hình thông qua các chỉ số như R-squared, MSE, MAE và phân tích biểu đồ giúp chúng ta hiểu rõ hơn về hiệu suất và những điểm cần cải thiện của mô hình.

Về mặt dữ liệu, chất lượng dữ liệu đầu vào ảnh hưởng trực tiếp đến kết quả dự đoán. Do đó, việc làm sạch dữ liệu, xử lý giá trị thiếu và ngoại lệ là rất quan trọng. Bên cạnh đó, việc lựa chọn các đặc trưng phù hợp và khai thác dữ liệu bằng các kỹ thuật như phân tích tương quan giúp tăng cường khả năng giải thích và dự đoán của mô hình.

Về mặt nghiệp vụ, sự hiểu biết sâu sắc về thị trường bất động sản và các yếu tố ảnh hưởng đến giá nhà là nền tảng để xây dựng mô hình hiệu quả và có giá trị thực tiễn. Việc giải thích rõ ràng kết quả dự đoán và các yếu tố tác động giúp người dùng hiểu rõ và tin tưởng vào mô hình hơn. Mô hình dự đoán giá nhà có thể ứng dụng rộng rãi trong các lĩnh vực như định giá bất động sản, tư vấn đầu tư, và quản lý tài sản, mang lại giá trị thực tế cho các bên liên quan.

Tóm lại, đồ án này không chỉ cung cấp một mô hình dự đoán giá nhà mà còn trang bị cho bọn em những kiến thức và kinh nghiệm quý báu về học máy, xử lý dữ liệu và ứng dụng vào bài toán thực tế. Qua đó, chúng ta có thể tiếp tục phát triển và hoàn thiện mô hình để đạt được hiệu suất dự đoán tốt hơn nữa trong tương lai.

.

4.2 - Những điều đã học được sau khi hoàn thành đồ án

- **Kiến thức:**

+ Chúng em đã biết như thế nào là việc cào dữ liệu từ một trang web bằng việc sử dụng thư viện Selenium, Selenium là một công cụ mạnh mẽ cho phép chúng ta tự động hóa các tác vụ trên trình duyệt web, nó cho phép chúng ta mô phỏng hành động của người dùng như nhấp chuột, nhập liệu, và điều hướng qua các trang web
+ Selenium cung cấp nhiều cách để tìm kiếm các phần tử trên trang web, trong đó phổ biến nhất là sử dụng CSS Selectors và XPath. CSS Selectors cho phép tìm kiếm phần tử dựa trên ID, class, hoặc các thuộc tính khác. XPath cung cấp cú pháp mạnh mẽ hơn để điều hướng qua cấu trúc của HTML.
+ Biết cách thức hoạt động của việc cào dữ liệu, cấu trúc của một trang web là như thế nào
+ Tìm tòi và học được cách xử lý một dữ liệu thô
+ Thành thạo về kiến thức code trong Python
+ Các kỹ thuật để triển khai một mô hình máy học cũng như xử lí các dữ liệu cần thiết cho một mô hình
+ Các kỹ năng đánh giá các mô hình hồi quy bằng các số liệu như MSE, MAE, R2-score
+ Khả năng trực quan hóa các dữ liệu thu được

- **Kỹ năng:**

+ Nâng cao khả năng làm việc nhóm, cùng nhau hoàn thành từng mục tiêu đã đặt ra từ đầu.
+ Gắn kết mọi người và nâng cao tinh thần đồng đội hơn với nhau
+ Tự tin nêu ra ý kiến và quan điểm cá nhân của mình để mọi người cùng nhau đánh giá và xem xét
+ Nâng cao khả năng tra cứu, đọc tài liệu và hệ thống chúng

4.3 – Hướng phát triển

4.3.1 - Cải thiện mô hình:

- Thử nghiệm các mô hình khác: Ngoài XGBoost và hồi quy tuyến tính, có thể thử nghiệm các mô hình khác như LightGBM, CatBoost.
- Tinh chỉnh siêu tham số nâng cao: Sử dụng các kỹ thuật như Grid Search, Random Search, hoặc Bayesian Optimization để tìm kiếm các giá trị siêu tham số tối ưu hơn.
- Kỹ thuật đặc trưng nâng cao: Áp dụng các kỹ thuật tạo đặc trưng nâng cao như Polynomial Features, Feature Interactions, hoặc sử dụng các mô hình nhúng từ (Word Embeddings) để biểu diễn các đặc trưng dạng văn bản (ví dụ: mô tả nhà).

4.3.2 - Mở rộng dữ liêu:

- Thu thập thêm dữ liệu: Cố gắng thu thập thêm dữ liệu từ nhiều nguồn khác nhau để làm phong phú thêm tập dữ liệu và giúp mô hình học tốt hơn.

- Bổ sung các đặc trưng mới: Tìm kiếm và thêm vào mô hình các đặc trưng mới có thể ảnh hưởng đến giá nhà, ví dụ như thông tin về hạ tầng xung quanh, khoảng cách đến các tiện ích công cộng, hoặc các chỉ số kinh tế xã hội.
- Sử dụng dữ liệu thời gian thực: Cập nhật dữ liệu thường xuyên để mô hình có thể phản ánh được những thay đổi mới nhất trên thị trường bất động sản.

4.3.3 - Triển khai ứng dụng thực tế:

- Xây dựng ứng dụng web/mobile: Phát triển một ứng dụng web hoặc ứng dụng di động cho phép người dùng nhập thông tin về căn nhà và nhận được dự đoán giá.
- Tích hợp với các hệ thống khác: Kết nối mô hình với các hệ thống khác như hệ thống quản lý bất động sản, hệ thống định giá, hoặc các nền tảng giao dịch bất động sản trực tuyến.
- Cung cấp API: Cung cấp API cho các ứng dụng khác có thể sử dụng để truy vấn và nhận dự đoán giá nhà.

4.3.4 - Nghiên cứu chuyên sâu:

- Phân tích yếu tố ảnh hưởng: Thực hiện các phân tích chuyên sâu để tìm hiểu rõ hơn về các yếu tố ảnh hưởng đến giá nhà và mức độ ảnh hưởng của từng yếu tố.
- Dự đoán xu hướng thị trường: Sử dụng mô hình để dự đoán xu hướng biến động giá nhà trong tương lai.
- Nghiên cứu các mô hình tiên tiến: Tìm hiểu và áp dụng các mô hình máy học tiên tiến như Deep Learning, Reinforcement Learning, hoặc các mô hình kết hợp để cải thiện hiệu suất dự đoán.

Các file tài liệu liên quan đến đồ án:

Daniz2k3/KHDL_Project (github.com)

TÀI LIỆU THAM KHẢO

- Dữ liệu tọa độ Việt Nam (https://cartographyvectors.com/map/1533-vietnam-with-regions?fbclid=IwZXh0bgNhZW0CMTAAAR00XqE1KArIDyrPrdvpYn7PjsiUCk3pmH0BSj8BYUg0nOYk72lupkniQSQ_aem_ZmFrZWR1bW15MTZieXRlcw#google_vignette)

- Cách sử dụng selenium : (Selenium with Python; XPath and CSS Selectors Guide)

(https://plotly.com/pythonapireference/generated/plotly.express.choropleth_mapbox.html)