XÂY DỰNG HỆ THỐNG DỰ ĐOÁN GIÁ TRỊ BẤT ĐỘNG SẢN

Nguyễn Thị Mai Trinh[1,2] and Nguyễn Thị Huyền Trang[1,2]

[1] Trường đại học Công nghệ Thông Tin Đại học quốc gia
[2] Đường Hàn Thuyên, Thủ Đức, Thành phố Hồ Chí Minh, Việt Nam {21522718, 21520488}@gm.uit.edu.vn

Tóm tắt nội dung Trong báo cáo này, nhóm tập trung vào việc xây dựng một hệ thống dự đoán giá trị bất động sản. Cụ thể, nhóm sử dụng một bộ dữ liệu tự thu thập, chứa đa dạng thông tin về giá bất động sản và các yếu tố liên quan. Với bộ dữ liệu này, nhòm thực hiện triển khai và đánh giá hiệu suất của các mô hình (Random Forest, Linear Regression, Isotonic Regression, Gradient Boosted, Factorization Machines Regression và Decision Tree) với các độ đo đánh giá phù hợp (RMSE, MAE, và R-squared (R2). Kết quả đạt được tốt ngoài mong đợi với độ chính xác R2 0.0782 và MAE 3,532.14 với mô hình Gradient Boosted. Do đó, nhóm tin rằng đề tài này sẽ góp phần mang đến cái nhìn rõ ràng hơn về quá trình định giá, giúp người tiêu dùng tự tin hơn khi quyết định mua bất động sản và hỗ trợ các chuyên gia định giá để nâng cao chất lượng dịch vụ và minh bạch trong giao dịch bất động sản. Đồng thời, nó cũng mở ra những cơ hội mới trong việc nghiên cứu và phát triển các phương pháp định giá tiên tiến trong ngành.

Keywords: Spark · bất động sản · Dự đoán giá nhà · Random Forest · Linear Regression · Isotonic Regression · Gradient Boosted

1 GIỚI THIỆU

Những nhà đầu tư bất động sản thường gặp phải khó khăn khi đối mặt với quyết định quan trọng: định giá căn hộ. Trong bối cảnh nền kinh tế khủng hoảng, thị trường bất động sản đình trệ, việc đưa ra một quyết định chính xác về giá trị của căn hộ trở nên ngày càng phức tạp và quan trọng.

Với mong muốn xây dựng một hệ thống dự đoán giá nhà, nhóm đã tự thu thập một bộ dữ liệu chứa đa dạng thông tin về giá bất động sản, loại bất động sản, địa chỉ, diện tích đất, cũng như thông tin chi tiết về người bán và người mua.

Trong quá trình phát triển, chúng tôi đã sử dụng các mô hình máy học trên nền tảng Spark ML, bao gồm: Random Forest, Linear Regression, Isotonic Regression, Gradient Boosted, Factorization Machines Regression và Decision Tree để thực hiện được với độ phức tạp và không đồng đều của dữ liệu.

Cấu trúc bài báo cáo như sau. Bộ dữ liệu và phương pháp xây dựng được trình bày trong phần II. Phần III mô tả ngắn gọn mô hình sử dụng, thử nghiệm

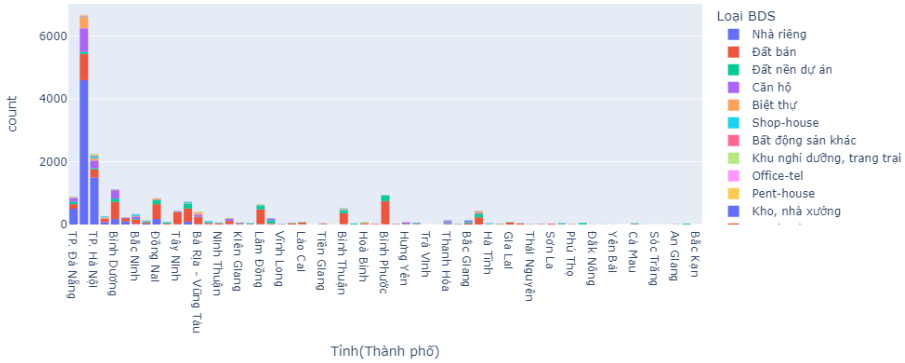
mô hình và trình bày kết quả. Kết quả thực nghiệm được trình bày trong phần IV của báo cáo. Cuối cùng, nhóm trình bày kết luận và hướng phát triển trong phần V của báo cáo.

2 BỘ DỮ LIỆU

2.1 Nguồn và quá trình thu thập

Bộ dữ liệu được chúng tôi thu thập một cách tự động từ sàn giao dịch môi giới bất động sản NhadatVui bằng kỹ thuật Web Scraping, sử dụng thư viện BeautifulSoup trong môi trường lập trình Python.

Bộ dữ liệu thô có nhiều dữ liệu bị khuyết thiếu, thuộc tính sai kiểu dữ liệu, còn chứa đơn vị ở một vài thuộc tính, cần phải tiền xử lý và làm sạch cẩn thận.

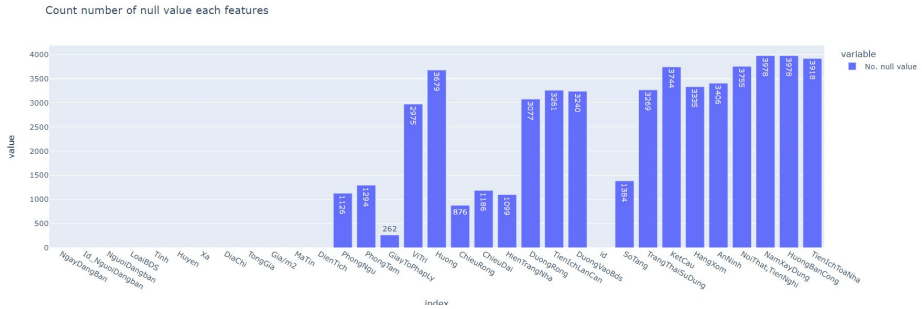


Hình 1. Biểu đồ Phân bố của dữ liệu theo loại bất động sản và tỉnh thành

Tập dữ liệu thu được là một tập hợp chồng chất các thông tin về bất động sản được đăng bán trên sàn, gồm 4,000 bài đăng, và mỗi bài đăng lại chứa đựng một lượng lớn thông tin chi tiết (33 thuộc tính). Bộ dữ liệu khá đa dạng và có khả năng đại diện trong việc mô phỏng thị trường bất động sản.

Trong tổng số hơn 30 thuộc tính, có 7 biến liên tục đo lường các giá trị có thể biến đổi một cách liên tục, 14 biến phân loại mô tả các đặc điểm có thể rời rạc được phân loại thành các nhóm, và 6 biến phân loại đa giá trị, là những biến có thể nhận một trong nhiều giá trị độc lập.

Ngoài ra, mỗi bản ghi còn chứa các biến quan trọng như 'MaTin' (biến định danh duy nhất cho từng bài đăng), 'NgayDangBan' (biến thời gian chỉ ra thời điểm đăng bán), và biến mục tiêu là 'TongGia' - đo lường tổng giá trị của bất động sản. Sự đa dạng về loại biến và tính chất của chúng tạo nên một thách thức đối với quá trình xử lý và mô hình hóa.



Hình 2. Số lượng biến null trên bộ dữ liệu ban đầu

Quá trình xử lý dữ liệu tiếp theo không chỉ đơn thuần là việc biến đổi dữ liệu thành định dạng phù hợp với mô hình máy học mà còn bao gồm việc xử lý dữ liệu, kiểm tra tính toàn vẹn, và chọn lọc các thuộc tính quan trọng. Tất cả những bước này là cơ sở cho sự thành công của quá trình nghiên cứu và đánh giá hiệu suất của các mô hình dự đoán giá trị bất động sản.

2.2 Tiền xử lý dữ liệu

Với bộ dữ liệu đã thu thập, nhóm thực hiện quá trình tiền xử lý và làm sạch, bao gồm năm bước quan trọng, mỗi bước được thiết kế để giải quyết một vấn đề cụ thể của dữ liệu đầu vào.

- Đồng bộ các giá trị Null và thay thế chúng: Các giá trị Null thường làm gián đoạn quá trình phân tích. Nếu có miền giá trị mang ý nghĩa tương tự, như các giá trị rỗng hoặc "–", chúng tôi đồng bộ hóa chúng để dễ dàng loại bỏ outlier khỏi bộ dữ liệu.
- Xoá bỏ đơn vị, đưa các biến về đúng kiểu dữ liệu: Việc này giúp tạo ra một tập dữ liệu đồng nhất và dễ xử lý hơn. Các biến đo lường, có thể chứa đơn vị không nhất quán, được chuẩn hóa để tránh những hiểu lầm không mong muốn trong quá trình phân tích. Cụ thể, nhóm cần loại bỏ đơn vị của các biến: 'TongGia', 'Gia/m2', 'ChieuDai', 'ChieuRong', 'DuongRong'. Trong đó, 2 biến 'TongGia', 'Gia/m2' ngoài việc loại bỏ đơn vị còn cần đưa chúng về độ đo đồng nhất. Cụ thể với biến 'TongGia', nhóm đưa cùng về đơn vị triệu, biến 'Gia/m2' được đưa về đơn vị đồng nhất là triệu/m2.
- Chuyển kiểu dữ liệu phù hợp: Sau khi loại bỏ đơn vị, đưa các biến 'TongGia', 'Gia/m2', 'ChieuDai', 'ChieuRong', 'DuongRong' từ kiểu string thành kiểu số float
- Xử lý dữ liệu khuyết: Chúng tôi áp dụng các phương pháp điền thay thế bằng giá trị trung bình (mean), giá trị xuất hiện nhiều nhất (mode),... Đối với các biến phân loại, nhóm sử dụng giá trị "Unknown" để điền khuyết dữ liệu khuyết nhằm đảm bảo không mất đi thông tin quan trọng mà giữ nguyên tính rời rạc của dữ liệu. Cụ thể: biến 'NamXayDung' được điền khuyết bởi

giá trị Mode, các biến 'PhongNgu' và 'PhongTam' được điền bởi gía trị mặc định là 0, biến 'SoTang' được điền mặc định là 1. Biến 'ChieuRong', 'DuongRong' được điền bởi giá trị mean. Các biến 'ChieuDai' sẽ được điền khuyết theo công thức sau để đảm bảo tính logic:

$$ChieuDai = DienTich/ChieuRong$$

– Loại bỏ dữ liệu bất hợp lý và cột hiếm xuất hiện và truyền kiểu: nhóm lựa chọn loại bỏ chúng để tăng tính độc lập và tính nhất quán của dữ liệu, tăng khả năng dự đoán của mô hình.

2.3 Rút trích đặc trưng

Dưới đây là những bước trích xuất đặc trưng được thực hiện trong bài báo cáo một cách kỹ thuật và chi tiết.

– Gom nhóm dữ liệu với Bucketizer: Trong bước này, chúng tôi sử dụng Bucketizer để phân cấp lại giá trị của biến phân loại 'Id-Nguoiban' dựa trên số lượng bất động sản bán được. Điều này giúp tạo ra một biến thay thế mới, giảm số lượng phân lớp; đồng thời tạo ý nghĩa cho biến này trong mô hình. Quá trình này giúp làm giảm chiều dài của vector đặc trưng và cải thiện khả năng tổng quát hóa của mô hình.
– Phân lớp tỉnh theo trình độ phát triển kinh tế: Đối với các thuộc tính 'Tinh', 'Huyen', 'Xa', nhóm sử dụng một bộ dữ liệu phụ chứa thông tin về trình độ phát triển kinh tế của từng tỉnh. Bằng cách này, nhóm phân loại lại biến 'Tinh' thành 4 cấp độ (Đặc biệt, I, II, III), phản ánh mức độ phát triển kinh tế của khu vực. Đối với biến 'Xa', 'Huyen', chúng tôi áp dụng phân cấp hành chính để gom nhóm chúng lại thành các loại địa giới phổ biến như 'Quan', 'Thi Xa', 'Huyen', 'Phuong', 'Thi Tran', 'Xa'.
– Sử dụng String Indexing và Onehot Encoding nhằm trích xuất thông tin từ các biến phân loại: giúp chuyển đổi các biến phân loại thành các vector encoded, tạo ra biểu diễn số học mà mô hình có thể sử dụng hiệu quả.
– Biến đổi dữ liệu Target ('TongGia') với Log Transformation: giúp giảm độ biến động của giá trị dự đoán 'TongGia', đồng thời đảm bảo rằng các giá trị dự đoán luôn là dương, giúp mô hình học được các mối quan hệ hiệu quả hơn.

2.4 Phân tích dữ liệu

Để đạt được kết quả tối ưu và khám phá nhiều giải pháp tiềm năng nhất, chúng tôi đã tích hợp nhiều phương pháp trong quá trình phân tích dữ liệu của mình. Trong đó, sử dụng dữ liệu outlier và chia nhóm dữ liệu là một phần quan trọng, nhằm tối ưu hóa hiệu suất của mô hình đối với bài toán cụ thể.

– Xử lý dữ liệu outlier: Xác định và loại bỏ những giá trị ngoại lệ có thể tạo ra nhiễu đối với mô hình. Thao tác này đảm bảo rằng mô hình sẽ được huấn luyện trên dữ liệu đồng nhất và không bị ảnh hưởng bởi các yếu tố ngoại lệ không mong muốn.

– Chia nhóm dữ: Áp dụng các kỹ thuật như cross-validation và holdout validation để đảm bảo tính chính xác và khả năng tổng quát của mô hình.

Dữ liệu sau khi trải qua quá trình xử lý và kết hợp với các kiến trúc thuật toán, trở nên phong phú và chứa đựng nhiều thông tin giá trị.

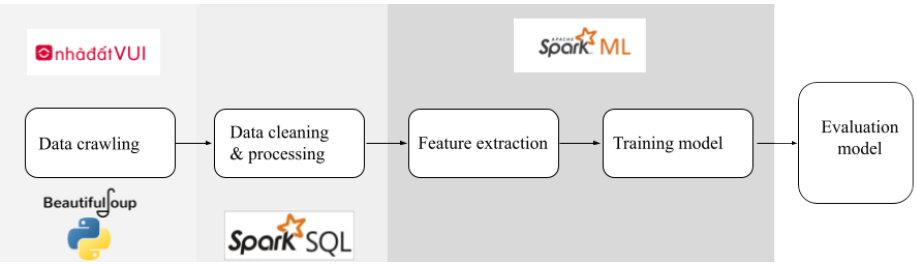
Quá trình này cũng giúp nhóm có thêm hiểu biết về đặc trưng của dữ liệu, hiểu rõ hơn về mối quan hệ phức tạp giữa các đặc trưng và biến mục tiêu, từ đó làm nền tảng cho việc tối ưu hóa và cải thiện mô hình trong tương lai.

3 THỰC NGHIỆM

3.1 Tổng quan quy trình

Sau khi thu thập và tiền xử lý bộ dữ liệu, ta được một tập dữ liệu đã được làm sạch chứa thông tin về giá trị bất động sản và các yếu tố phong phú và chính xác. Sau đó, nhóm sử dụng các mô hình máy học trên nền Spark ML, bao gồm: Linear Regression, Decision Tree, Random Forest, Isotonic Regression, và Gradient Boosting để thực nghiệm trên bộ dữ liệu nhằm dự đoán giá trị bất động sản. Mỗi mô hình được điều chỉnh và tối ưu để dự đoán giá trị của bất động sản một cách chính xác và hiệu quả. Sau đó, nhóm tiến hành phân tích đánh giá hiệu suất dự đoán để lựa chọn ra những mô hình xuất sắc và phù hợp nhất bộ dữ liệu.

Quy trình xây dựng mô hình được thể hiện trong Hình 3



Hình 3. Quy trình thực hiện

3.2 Các phương pháp máy học

Nhóm đã tận dụng sức mạnh của thư viện Pyspark để tải mô hình và kiến trúc thuật toán, giúp giảm bớt khối lượng công việc và mang lại sự đa dạng và sức mạnh trong quá trình triển khai mô hình. Cụ thể nhóm sử dụng 5 mô hình có sẵn:

Thuật toán Decision Tree - cây quyết định là một trong những mô hình có khả năng diễn giải cao và có thể thực hiện cả nhiệm vụ classification và regression.

Với Decision Tree, chúng ta không cần phải thực hiện bất kỳ loại xử lý trước dữ liệu nào trước đó. Decision Tree đủ mạnh để xử lý tất cả các loại vấn đề như giá trị ngoại lệ, giá trị bị thiếu, đa cộng tuyến. Bên cạnh đó, Decision Tree có khả năng xử lý dữ liệu phi tuyến mà các mô hình tuyến tính cổ điển không xử lý được.

Thuật toán Random Forest hay "Rừng Ngẫu Nhiên" là một phương pháp kết hợp nhiều cây quyết định (Decision Trees) để tạo ra một mô hình mạnh mẽ hơn và có khả năng tổng quát hóa cao hơn.

Thuật toán Random Forest bắt đầu bằng việc chọn ngẫu nhiên một tập con của dữ liệu từ tập huấn luyện. Tiếp theo, nó xây dựng một cây quyết định trên tập con này. Quá trình này được lặp lại nhiều lần để tạo ra nhiều cây quyết định khác nhau.

Thuật toán Linear Regression là một trong những thuật toán cơ bản và phổ biến nhất của Supervised Learning, trong đó đầu ra dự đoán là liên tục. Thuật toán này thích hợp để dự đoán các giá trị đầu ra là các đại lượng liên tục như doanh số hay giá cả,...

Thuật toán Gradient Boosting là một phương pháp học máy tiên tiến được thiết kế để xây dựng mô hình dự đoán mạnh mẽ bằng cách tận dụng sức mạnh của nhiều mô hình yếu. Trong lớp các thuật toán học máy, Gradient Boosting nổi bật về hiểu và khai thác sự mạnh mẽ của dữ liệu.

Thuật toán Isotonic Regression là một công cụ hữu ích, chủ yếu được sử dụng để mô hình hóa các quan hệ đơn điệu và không giả định hình dạng của đường chính.

Mục tiêu của Isotonic Regression là tìm ra một hàm số monotonically increasing hoặc monotonically decreasing sao cho sai số giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất.

3.3 Độ đo đánh giá

Để đánh giá hiệu suất của mô hình dự đoán giá nhà, nhóm sử dụng các độ đo phù hợp với bài toán như RMSE, MAE, và R^2. Các độ đo này cung cấp cái nhìn toàn diện về khả năng dự đoán của mô hình và mức độ chính xác của nó.

Root Mean Square Error (RMSE) là một thước đo sự chênh lệch giữa giá trị dự đoán và giá trị thực tế của mục tiêu được tính bằng công thức:

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

Mean Absolute Error (MAE) là một độ đo đơn giản của sự chênh lệch trung bình giữa giá trị dự đoán và giá trị thực tế, được tính như sau:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Coefficient of Determination ($\mathbf{R^2}$) là một độ đo thống kê đánh giá khả năng của mô hình so với một mô hình đơn giản như mô hình trung bình. Công thức tính ^{2} là:

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

Trong các công thức trên, các biểu thức được giải thích như sau:

- n là số lượng mẫu dữ liệu.
- y_i là giá trị thực tế của mẫu thứ i.
- $\hat{y}_i$ là giá trị dự đoán tương ứng với mẫu thứ i.
- $\bar{y}$ là giá trị trung bình của các giá trị thực tế y.

4 Kết quả thực nghiệm

Thực nghiệm của nhóm đưa ra cái nhìn tổng quan về hiệu suất của mô hình khi dự đoán trên bộ dữ liệu đã thu thập. Với kết quả thực nghiệm, nhóm nhận định Gradient Boosted là mô hình có kết quả cao nhất để giải quyết bài toán.

Mô hình	RMSE	MAE	R^2
Linear regression	$194*10^9$	$4*10^9$	$-69*10^{12}$
Isotonic regression	23594.512	5541.002	-0.0222
Decision tree	22,724.981	4,052.877	0.0550
Random forest	22,403.987	3,692.714	0.0516
Gradient boosted	11,343.9	3,532.14	0.0782

Bảng 1. Đánh giá các phương pháp trên tập kiểm thử

4.1 Nhận xét

Kết quả thu được từ việc triển khai các mô hình được đánh giá một cách kỹ lưỡng và so sánh để xác định mức độ hiệu quả của từng thuật toán. Quá trình này bao gồm việc đánh giá hiệu suất dự đoán, xác định độ chính xác và độ nhạy của mỗi mô hình, cũng như đánh giá khả năng xử lý dữ liệu thực tế và độ ổn định của thuật toán. Từ đó, nhóm có các nhận xét sau:

- Mô hình Linear Regression cho ra kết quả không khả quan với RMSE và MAE vô cùng cao, cho thấy sự chênh lệch lớn giữa giá trị dự đoán và giá trị thực tế. Giá trị R^2 âm và vô cùng lớn, ngụ ý rằng mô hình không thể giải thích được phần lớn sự biến động của dữ liệu và không có tính ứng dụng để dự đoán.

- Mô hình Isotonic Regression: mặc dù RMSE và MAE đã giảm đáng kể so với Linear Regression, nhưng kết quả chưa đủ tốt. Giá trị R^2 gần bằng 0 và âm, ngụ ý rằng mô hình không thể giải thích sự biến động của dữ liệu, dự đoán không chính xác và không đáng tin cậy.

- Mô hình Decision Tree: mặc dù có RMSE và MAE cao, nhưng đã có một bước tiến với R^2 dương, cho thấy mô hình có thể giải thích được một phần nào đó của sự biến động trong dữ liệu.

- Mô hình Random Forest cũng có kết quả tương tự như Decision Tree, được cải thiện nhỏ về RMSE và MAE.

- Gradient Boosted là mô hình có kết quả tốt nhất trong số các mô hình được đánh giá, với RMSE và MAE thấp hơn và R^2 dương, điều này nói lên rằng mô hình giải thích được một phần nhất định của sự biến động trong dữ liệu.

Tóm lại, các mô hình Decision Tree, Random Forest và Gradient Boosted cho kết quả tốt hơn so với Linear Regression và Isotonic Regression. Trong đó, tuy mô hình Grandient boosted cho kết quả dự đoán tốt hơn các mô hình khác, nhưng độ đo MAE = 3,532.14 hay RMSE = 11,343.9 vẫn khá cao khi đơn vị của biến 'TongGia' là triệu đồng. Vì vậy cần cải thiện bộ dữ liệu và các mô hình để kết quả dự đoán đáng tin cậy và có ý nghĩa hơn.

5 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong phạm vi báo cáo này, nhóm đã tiến hành nghiên cứu chi tiết và triển khai một hệ thống dự đoán giá trị bất động sản theo thời gian thực. Dữ liệu được chúng tôi thu thập từ nguồn nhadatvui.vn, được xử lý và làm sạch một cách cẩn thận, tạo ra một bộ dữ liệu đầy đủ và độc đáo.

Qua quá trình thử nghiệm trên nhiều thuật toán khác nhau với sự hỗ trợ của PySpark MLlib, nhóm nhận định: Gradient Boosted đã nổi bật cho ra kết quả dự đoán tốt nhất với độ đo RMSE đạt 11,343.9, MAE đạt 3,532.14, R2 chỉ 0.0782.

Qua đề tài này, nhóm hiểu rõ hơn về mối quan hệ phức tạp giữa các đặc trưng và biến mục tiêu, từ đó làm nền tảng cho việc tối ưu hóa và cải thiện mô hình trong tương lai.

Trong tương lai, chúng tôi đã đặt ra những mục tiêu cụ thể để cải thiện hệ thống, bao gồm việc nghiên cứu và áp dụng các phương pháp kết hợp mới để tăng hiệu suất, rút trích thông tin từ biến có tương quan thấp để làm giàu dữ liệu, và mở rộng nguồn dữ liệu thông qua việc thu thập thêm nhiều nguồn dữ liệu đa dạng. Những bước tiến này sẽ giúp hệ thống ngày càng hoàn thiện, đồng thời đáp ứng mạnh mẽ hơn với sự biến động của thị trường bất động sản.

Tài liệu

[1] Ch. Raga Madhuri, G Anuradha, and M. Vani Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," Mar. 2019, doi: https://doi.org/10.1109/icsss.2019.8882834.

[2] X. Meng et al., "MLlib: Machine Learning in Apache Spark," Journal of Machine Learning Research, vol. 17, pp. 1–7, 2016, Available: https://www.jmlr.org/papers/volume17/15-237/15-237.pdf

[3] S. Salloum, Ruslan Dautov, X. Chen, Patrick Xiaogang Peng, and Joshua Zhexue Huang, "Big data analytics on Apache Spark," International Journal of Data Science and Analytics, vol. 1, no. 3–4, pp. 145–164, Oct. 2016, doi: https://doi.org/10.1007/s41060-016-0027-9.

[4] S. Sanyal, Saroj Kumar Biswas, D. Das, M. Chakraborty, and Biswajit Purkayastha, "Boston House Price Prediction Using Regression Models," 2022 2nd International Conference on Intelligent Technologies (CONIT), Jun. 2022, doi: https://doi.org/10.1109/conit55038.2022.9848309.

[5] [7] "Understanding from Machine Learning Models | The British Journal for the Philosophy of Science: Vol 73, No 1," The British Journal for the Philosophy of Science, 2022. [Trực tuyến] Link: https://www.journals.uchicago.edu/doi/abs/10.1093/bjps/axz035?journalCode=bjps (accessed Jan. 03, 2024).

[6] A. J. Bency, Swati Rallapalli, R. Ganti, Mudhakar Srivatsa, and B. S. Manjunath, "Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery," eScholarship University of California (University of California), Mar. 2017, doi: https://doi.org/10.1109/wacv.2017.42. ∎

[7] Abhaya Abhaya and Binoy Krishna Patra, "RDPOD: an unsupervised approach for outlier detection," Neural Computing and Applications, vol. 34, no. 2, pp. 1065–1077, Aug. 2021, doi: https://doi.org/10.1007/s00521-021-06432-6.

[8] P. R. Selvin, A. Maheshwari, and Prashant Johri, "Comparative Analysis of Ml Algorithms Stream Lit Web Application", 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Dec. 2021, doi: https://doi.org/10.1109/icac3n53548.2021.9725496.

[9] Nhà đất vui, [Trực tuyến]. Link: https://nhadatvui.vn/mua-ban-nha-dat

[10] G. Varoquaux and Olivier Colliot, "Evaluating Machine Learning Models and Their Diagnostic Value," Neuromethods, pp. 601–630, Jan. 2023, doi: https://doi.org/10.1007/978-1-0716-3195-9$_2$0.