**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TIỂU LUẬN CHUYÊN ĐỀ TỐT NGHIỆP KHDL 1**

**XÂY DỰNG MÔ HÌNH DỰ ĐOÁN GIÁ NHÀ SỬ DỤNG MACHINE LEARNING TRONG PYTHON**

**Giảng viên hướng dẫn : TS. VÕ THỊ HỒNG THẮM**

**Sinh viên thực hiện : VÕ THÁI HIỂN**

**MSSV : 2000005690**

**Khoá : 2020**

**Ngành : CÔNG NGHỆ THÔNG TIN**

**Chuyên ngành : KHOA HỌC DỮ LIỆU**

TpHCM, tháng 12 năm 2023

**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TIỂU LUẬN CHUYÊN ĐỀ TỐT NGHIỆP KHDL 1**

**XÂY DỰNG MÔ HÌNH DỰ ĐOÁN GIÁ NHÀ SỬ DỤNG MACHINE LEARNING TRONG PYTHON**

**Giảng viên hướng dẫn : TS. VÕ THỊ HỒNG THẮM**

**Sinh viên thực hiện : VÕ THÁI HIỂN**

**MSSV : 2000005690**

**Khoá : 2020**

**Ngành : CÔNG NGHỆ THÔNG TIN**

**Chuyên ngành : KHOA HỌC DỮ LIỆU**

TpHCM, tháng 12 năm 2023

LỜI MỞ ĐẦU

Sự phát triển mạnh mẽacủa cuộc cách mạng khoa họcavà công nghệ, đặc biệt là của công nghệathông tin và truyền thông đã tác động sâu sắc đến sự phát triển kinhatế thế giới. Thế giới đang chuyểnadần từ xã hội công nghiệp sang xã hội thông tin với nềnakinh tế chi thức. Toàn cầuahóa kinh tế được xác định là một xu thế khách quan, tạo cơahội cho phát triển. Khoa họcavà công nghệ (KH&CN) sẽ cóabước nhảy vọt và những đột phá lớn, tác độnganhều mặt đến tất cả cácaquốc gia

Những năm gầnađây, vai trò của các hệ thốngathông tin trong doanh nghiệp đã ngày càng lớn mạnh. Từ chỗachỉ được sử dụng để hỗ trợ một số hoạtađộng trong văn phòng, hệ thống thôngatin đã trở nên có vai trò chiến lược trongadoanh nghiệp. Đặc biệt nhữngathành tựu trong công nghệathông tin (CNTT) đã khiến doanh nghiệp ngày càngachú ý hơn tới việc áp dụng những nó đểagia tăng thế cạnh tranh và tạo cơahội cho mình. Thành tựu CNTT không chỉ giới hạnatrong các doanh nghiệp lớn, tầm cỡ đa quốc gia màacòn lan rộng trong tất cả các doanh nghiệp, kể cả những doanhanghiệp vừa và nhỏ ở những nướcađang phát triển. Tuy nhiên, việc ứng dụngathành tựu CNTT một ứng dụng thành công trongacác lĩnh vực hiện nay cũngacòn mới ở nước ta. Vì vậy, đồ án của nhóm đó là tìmahiểu về “Xây dựng mô hình dự đoán giá nhà sử dụng Machine learning trong python” để có cái nhìn sâu hơn, thực tế hơn vềaviệc ứng dụng hệ thống thông tin trongacác lĩnh vực đời sống. Vìathời gian thực hiện không nhiềuanên chắc chắn còn nhiều thiếu sót, mong được thầyacô chỉ dẫn thêm

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành đến Cô về môn học “Chuyên đề tốt nghiệp KHDL 1” trong học kỳ vừa qua tại Trường Đại học Nguyễn Tất Thành. Cô đã tạo ra một môi trường học tập đầy thú vị và bổ ích, giúp em hiểu rõ hơn về lĩnh vực quan trọng này.

Em thực sự trân trọng cách Cô trình bày kiến thức một cách rõ ràng, dễ hiểu và thú vị. Nhờ vào sự hướng dẫn tận tâm của Cô, em đã có cơ hội tiếp cận những khái niệm phức tạp một cách dễ dàng hơn. Những ví dụ thực tế và bài tập thực hành trong khóa học đã giúp em áp dụng kiến thức vào thực tế và phát triển kỹ năng thực sự cần thiết.

Không chỉ giới hạn trong việc truyền đạt kiến thức, Cô còn tạo điều kiện cho chúng em thảo luận, trao đổi ý kiến và học hỏi từ nhau. Điều này thực sự đã tạo ra một không gian học tập tích cực và khuyến khích sự phát triển cá nhân của từng học viên.

Em cảm ơn Cô vì sự kiên nhẫn và lòng nhiệt tình trong việc giải đáp các thắc mắc của em. Dù là trong lớp học hay ngoài giờ, Cô luôn sẵn sàng hỗ trợ và động viên em vượt qua những khó khăn trong quá trình học tập.

Cuối cùng, em xin chân thành cảm ơn Cô vì sự dạy dỗ tận tâm và những kiến thức bổ ích mà Cô đã truyền đạt cho chúng em. Môn học này không chỉ giúp em nắm vững kiến thức về Khoa học dữ liệu, mà còn giúp em phát triển tư duy phân tích, giải quyết vấn đề và học hỏi cách tiếp cận các thách thức trong tương lai.

Một lần nữa, em xin bày tỏ lòng biết ơn chân thành đến Cô Võ Thị Hồng Thắm. Mong rằng Cô sẽ tiếp tục có những đóng góp quý báu trong việc truyền đạt kiến thức cho thế hệ học viên tương lai.

Em xin chân thành cảm ơn!

**Võ Thái Hiển**

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ 2 NĂM HỌC 2022 – 2023** |

**PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN**

BM-ChT-11

Môn thi: Đồ án Chuyên ngành Khoa học dữ liệu Lớp học phần: 20DTH1D

Nhóm sinh viên thực hiện:

1. Châu Hùng Anh Tham gia đóng góp: 100%

2. Nguyễn Tiến Thành Tham gia đóng góp: 100%

3. Võ Thái Hiển Tham gia đóng góp: 100%

3. Nguyễn Minh Đăng Tham gia đóng góp: 100%

Ngày thi: Phòng thi:

Đề tài tiểu luận/báo cáo của sinh viên : Xây dựng mô hình dự đoán giá nhà sử dụng Machine learning trong python

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo |  |  |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần |  |  |  |
| * Lập luận |  |  |  |
| * Kết luận |  |  |  |
| Trình bày |  |  |  |
| **TỔNG ĐIỂM** |  |  |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)* |

**Võ Thị Hồng Thắm**

MỤC LỤC

[LỜI MỞ ĐẦU i](#_Toc152964809)

[LỜI CẢM ƠN ii](#_Toc152964810)

[MỤC LỤC iv](#_Toc152964811)

[DANH MỤC HÌNH vi](#_Toc152964812)

[KÝ HIỆU CÁC CỤM TỪ VIẾT TẮT vii](#_Toc152964813)

[CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI 8](#_Toc152964814)

[1.1. Lý do chọn đề tài 8](#_Toc152964815)

[1.2. Mục đích nghiên cứu 8](#_Toc152964816)

[1.3. Đối tượng nghiên cứu 9](#_Toc152964817)

[1.4. Phạm vi nghiên cứu 9](#_Toc152964818)

[1.5. Phương pháp nghiên cứu 9](#_Toc152964819)

[1.6. Ý nghĩa và ứng dụng 10](#_Toc152964820)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 11](#_Toc152964821)

[2.1. Tổng quan về giá nhà 11](#_Toc152964822)

[2.1.1. Nhận định về giá nhà 11](#_Toc152964823)

[2.1.2. Thách thức cho người mua nhà 11](#_Toc152964824)

[2.2. Tổng quan về Machine Learning (Học máy) 12](#_Toc152964825)

[2.2.1. Tổng quan về Supervised Learning (Học có giám sát) 12](#_Toc152964826)

[2.2.2. Tổng quan về Unsupervised Learning (Học không giám sát) 13](#_Toc152964827)

[2.2.3. Tổng quan về Reinforcement Learning (Học tăng cường) 14](#_Toc152964828)

[2.3. Tổng quan về Regression trong Machine Learning 15](#_Toc152964829)

[2.3.1. Linear Regression (Hồi Quy Tuyến Tính) 15](#_Toc152964830)

[2.3.2. Random Forest Regression (Rừng Ngẫu Nhiên Hồi Quy) 16](#_Toc152964831)

[2.4. Support Vector Machine (SVM) 17](#_Toc152964832)

[2.4.1. Cơ Bản về Support Vector Machine 17](#_Toc152964833)

[2.4.2. Các Khái Niệm Quan Trọng trong SVM 17](#_Toc152964834)

[2.4.3. Ứng Dụng và Ưu Điểm 18](#_Toc152964835)

[CHƯƠNG 3. MÔ HÌNH THỰC NGHIỆM 20](#_Toc152964836)

[3.1. Quy trình thực hiện mô hình 20](#_Toc152964837)

[3.2. Các kỹ thuật chuẩn bị dữ liệu sử dụng 21](#_Toc152964838)

[3.3. Tông quan bộ dữ liệu 22](#_Toc152964839)

[3.4. Thực nghiệm mô hình 22](#_Toc152964840)

[3.5. Kết quả đạt được 38](#_Toc152964841)

[CHƯƠNG 4. KẾT LUẬN 40](#_Toc152964842)

[4.1. Kết luận 40](#_Toc152964843)

[4.2. Hướng phát triển 40](#_Toc152964844)

[TÀI LIỆU THAM KHẢO 41](#_Toc152964845)

DANH MỤC HÌNH

[Hình 3. 1: Quy trình thực hiện mô hình 19](#_Toc152964199)

[Hình 3. 2: Nhập các thư viện và tập dữ liệu 22](#_Toc152964200)

[Hình 3. 3: Thông tin và kích thước của tập dữ liệu 22](#_Toc152964201)

[Hình 3. 4: Ví dụ code phân loại các đặc trưng 23](#_Toc152964202)

[Hình 3. 5: Tương quan của các đặc trưng 24](#_Toc152964203)

[Hình 3. 6: Các giá trị unique của các đặc trưng Categorical 24](#_Toc152964204)

[Hình 3. 7: Số lượng thực tế của các đặc trưng Categorical 25](#_Toc152964205)

[Hình 3. 8: Ví dụ code làm sạch dữ liệu 25](#_Toc152964206)

[Hình 3. 9: Dữ liệu trước khi xử lý 26](#_Toc152964207)

[Hình 3. 10: Dữ liệu sau khi xử lý 26](#_Toc152964208)

[Hình 3. 11: Ví dụ code xóa các dòng bị thiếu và kiểm tra dữ liệu còn thiếu không 27](#_Toc152964209)

[Hình 3. 12: Ví dụ code sử dụng OneHotEncode 27](#_Toc152964210)

[Hình 3. 13: OneHotEncoding cho toàn bộ danh sách 28](#_Toc152964211)

[Hình 3. 14: Ví dụ code chia tập dữ liệu thành train và test 28](#_Toc152964212)

[Hình 3. 15: Công thức tính MAPE 30](#_Toc152964213)

[Hình 3. 16: Công thức MSE 30](#_Toc152964214)

[Hình 3. 17: Công thức MRSE 31](#_Toc152964215)

[Hình 3. 18: Chỉ số đánh giá các mô hình 31](#_Toc152964216)

[Hình 3. 19: RMSE Score của từng mô hình 32](#_Toc152964217)

[Hình 3. 20: thay thế OneHotEncoder bằng OrdinalEncoder 32](#_Toc152964218)

[Hình 3. 21: Độ quan trọng của các đặc trưng 33](#_Toc152964219)

[Hình 3. 22: Sử dụng StandardScaler để chuẩn hóa dữ liệu 33](#_Toc152964220)

[Hình 3. 23: Chỉ số đánh giá các mô hình sau khi thay đổi, thêm các kỹ thuật 34](#_Toc152964221)

[Hình 3. 24: RMSE Score của từng mô hình sau khi thay đổi, thêm các kỹ thuật 34](#_Toc152964222)

KÝ HIỆU CÁC CỤM TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| Chữ viết tắt | English | Tiếng việt |
| CSV | Comma-Separated Values | Giá trị được Phân tách bằng Dấu Phẩy |
| MAPE | Mean Absolute Percentage Error | Sai số Tuyệt đối Trung bình Theo Phần trăm |
| MSE | Mean Squared Error | Sai số Bình phương Trung bình. |
| RMSE | Root Mean Squared Error | Sai số Bình phương Trung bình Căn bậc hai |
| SVC | Support Vector Classification | Phân loại Vector Hỗ trợ |
| SVM | Support Vector Machine | Máy Vector Hỗ trợ |

CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI

1.1. Lý do chọn đề tài

Dự án xây dựng mô hình dự đoán giá nhà sử dụng machine learning trong Python là một lựa chọn sáng tạo và hứa hẹn. Việc dự đoán giá nhà là một vấn đề thực tế và có ứng dụng rộng rãi trong thị trường bất động sản. Nếu mô hình của bạn hiệu quả, nó có thể giúp những người mua và bán nhà đưa ra quyết định thông tin.

Dự án này cung cấp cơ hội để làm việc với dữ liệu đa dạng, bao gồm thông tin về diện tích, vị trí, tiện ích xung quanh, và nhiều yếu tố khác có thể ảnh hưởng đến giá nhà. Machine learning mang lại khả năng dự đoán tốt hơn khi có nhiều biến và mối quan hệ phức tạp. Việc áp dụng các thuật toán machine learning như regression, decision trees, hoặc neural networks có thể giúp tạo ra mô hình phức tạp hơn so với các phương pháp truyền thống. Dự án này cung cấp cơ hội để học và áp dụng kiến thức về Python và machine learning trong một dự án thực tế. Điều này có thể giúp bạn phát triển kỹ năng và hiểu biết sâu rộng về lĩnh vực này.

Việc đối mặt với thách thức của việc xây dựng một mô hình dự đoán giá nhà có thể kích thích sự sáng tạo và khám phá trong quá trình học.

1.2. Mục đích nghiên cứu

Nghiên cứu này có thể giúp mình hiểu rõ hơn về yếu tố nào ảnh hưởng đến giá nhà và làm thế nào chúng tương tác với nhau trong thị trường bất động sản.

Dự án giúp bạn áp dụng kiến thức về machine learning vào một vấn đề thực tế. Việc này giúp củng cố kỹ năng thực hành và hiểu biết về cách machine learning có thể được tích hợp vào các ứng dụng thực tế.

Dự án này có thể là cơ hội để tìm hiểu về quy trình phân tích dữ liệu từ việc thu thập và làm sạch dữ liệu đến xây dựng và đánh giá mô hình. Điều này có thể mang lại hiểu biết sâu rộng về cách tiếp cận và giải quyết vấn đề trong lĩnh vực phân tích dữ liệu.

Dự án này có thể là cơ hội để tìm hiểu về quy trình phân tích dữ liệu từ việc thu thập và làm sạch dữ liệu đến xây dựng và đánh giá mô hình. Điều này có thể mang lại hiểu biết sâu rộng về cách tiếp cận và giải quyết vấn đề trong lĩnh vực phân tích dữ liệu. Tóm lại mục đích nghiên cứu sự phát triển cá nhân, hiểu biết sâu rộng về thị trường bất động sản, và đóng góp vào lĩnh vực machine learning và phân tích dữ liệu.

1.3. Đối tượng nghiên cứu

Các doanh nghiệp hoặc tổ chức trong lĩnh vực bất động sản có thể tận dụng thông tin từ mô hình để hiểu thêm về thị trường và đưa ra quyết định chiến lược.

Đối tượng nghiên cứu không chỉ giới hạn trong cộng đồng machine learning mà còn mở rộng đến những người liên quan trực tiếp đến thị trường bất động sản và quyết định mua bán nhà.

1.4. Phạm vi nghiên cứu

Xác định các yếu tố ảnh hưởng đến giá nhà mà mô hình sẽ sử dụng để đưa ra dự đoán. Điều này có thể bao gồm diện tích, vị trí địa lý, số phòng, tiện ích xung quanh, và các yếu tố khác có thể ảnh hưởng đến giá nhà.

Xác định khoảng thời gian mà nghiên cứu sẽ tập trung vào. Có thể là một khoảng thời gian cụ thể hoặc nghiên cứu có thể áp dụng cho nhiều giai đoạn thời kỳ khác nhau.

Xác định ngôn ngữ lập trình (Python) và các thư viện machine learning cụ thể (scikit-learn, TensorFlow, PyTorch) sẽ được sử dụng trong dự án.

Phạm vi nghiên cứu giúp định hình rõ ràng cho dự án, đồng thời giúp đảm bảo rằng nghiên cứu được thực hiện một cách có tổ chức và có tính ứng dụng cao.

1.5. Phương pháp nghiên cứu

Xác định và thu thập dữ liệu liên quan đến giá nhà, bao gồm thông tin về diện tích, vị trí, tiện ích xung quanh, số phòng, v.v.

Lựa chọn một hoặc nhiều mô hình machine learning phù hợp cho bài toán dự đoán giá nhà. Các lựa chọn có thể bao gồm Linear Regression, Decision Trees, Random Forests, hoặc các mô hình deep learning nếu cần thiết.

Sử dụng tập kiểm thử để đánh giá hiệu suất của mô hình. Các metric đánh giá như Mean Squared Error (MSE), R-squared, hay các metric khác có thể được sử dụng.

Duyệt kết quả và xuất báo cáo về mô hình, bao gồm cả các kết quả, giải thích về quyết định của mô hình, và cách sử dụng nó.

Phương pháp này nhấn mạnh quy trình từ thu thập dữ liệu đến triển khai mô hình và chia sẻ kết quả, giúp đảm bảo tính toàn vẹn và hiệu quả của dự án.

1.6. Ý nghĩa và ứng dụng

Dự án cung cấp thông tin hữu ích cho người mua và bán nhà, giúp họ đưa ra quyết định mua bán dựa trên dự đoán giá nhà chính xác và thông tin chi tiết về các yếu tố ảnh hưởng.Chủ sở hữu bất động sản có thể sử dụng mô hình để định giá tài sản của họ một cách chính xác hơn, giúp họ tối ưu hóa giá và tăng khả năng bán nhanh chóng.

Các doanh nghiệp và tổ chức trong ngành bất động sản có thể tích hợp mô hình vào các hệ thống của họ để hỗ trợ quyết định và cung cấp dịch vụ chính xác hơn cho khách hàng. Dự án mang lại cơ hội để phát triển kỹ năng trong lĩnh vực machine learning và sử dụng ngôn ngữ lập trình Python trong môi trường thực tế. Tóm lại, dự án này không chỉ mang lại lợi ích trong lĩnh vực bất động sản mà còn có ý nghĩa rộng lớn trong việc áp dụng machine learning và Python vào thực tế.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. Tổng quan về giá nhà

Giá nhà, một trong những yếu tố quyết định cuộc sống của chúng ta, ngày càng trở nên ngày càng khó khăn đối với người mua nhà. Trong bối cảnh thị trường bất động sản đang phát triển nhanh chóng và với nhiều thách thức kinh tế, chúng ta cần xem xét cẩn thận vấn đề này để đảm bảo tính bền vững của thị trường và cung cấp cơ hội mua nhà công bằng cho mọi người.

2.1.1. Nhận định về giá nhà

**Tăng Trưởng Kinh Tế và Lãi Suất:** Dù sự tăng trưởng kinh tế thường kèm theo sự tăng cầu mua nhà, nhưng lãi suất thị trường biến động đang tạo ra một bức tranh không chắc chắn. Các người mua nhà đang phải đối mặt với sự không chắc chắn về khả năng thanh toán trong tương lai.

**Vị Trí và Địa Lý:** Sự chênh lệch về giá nhà giữa các khu vực đang ngày càng lớn. Những người muốn định cư gần trung tâm đô thị và tiện ích đang phải đối mặt với giá nhà không dễ dàng tiếp cận.

**Chính Sách Tài Chính Quốc Gia:** Chính sách tài chính quốc gia và chính sách thuế bất động sản đang góp phần vào sự không chắc chắn và áp lực về giá nhà.

**Ảnh Hưởng Của Công Nghệ:** Sự phát triển của công nghệ và xu hướng làm việc từ xa đang thay đổi cách mọi người đánh giá vị trí cần sống, có thể tạo ra sự chuyển động lớn trong thị trường bất động sản.

**Chấn Thương Tâm Lý của Người Mua:** Sự lo ngại về giá nhà có thể tạo ra những ảnh hưởng tiêu cực về mặt tâm lý, khiến người mua đua đòi và quyết định mua nhà mà họ có thể không hoàn toàn chấp nhận được.

2.1.2. Thách thức cho người mua nhà

**Tăng Giá Liên Tục:** Sự tăng giá liên tục tạo ra một tâm lý "sợ lỡ cơ hội" cho người mua nhà. Nhiều gia đình đang phải cân nhắc giữa việc mua ngay và chờ đợi một sự ổn định giá.

**Chênh Lệch Cung và Cầu:** Tình trạng chênh lệch giữa cung và cầu tạo ra một thị trường không công bằng, khiến người mua nhà phải cạnh tranh khốc liệt và thậm chí làm tăng giá.

**Chính Sách Tài Chính Ngân Hàng:** Người mua nhà, đặc biệt là nhóm thu nhập thấp, đang phải đối mặt với những yêu cầu về điều kiện vay khắt khe và áp lực tăng lãi suất.

2.2. Tổng quan về Machine Learning (Học máy)

Machine Learning (Học máy) là một lĩnh vực trong ngành trí tuệ nhân tạo (Artificial Intelligence) mà trong đó máy tính được huấn luyện để học từ dữ liệu và tự động cải thiện hiệu suất theo thời gian. Mục tiêu chính của Machine Learning là phát triển các thuật toán và mô hình có khả năng nhận dạng mẫu, tìm hiểu và thực hiện các dự đoán hoặc hành động mà không cần được lập trình cụ thể.

Machine Learning có thể được chia thành ba loại chính: supervised learning (học có giám sát), unsupervised learning (học không giám sát) và reinforcement learning (học củng cố). Trong supervised learning, mô hình được đào tạo bằng cách sử dụng một tập dữ liệu đã được gán nhãn. Trong unsupervised learning, mô hình học từ dữ liệu không được gắn nhãn. Còn trong reinforcement learning, mô hình học thông qua việc tương tác và nhận phản hồi từ môi trường xung quanh.

Các ứng dụng của Machine Learning rất đa dạng và đang được áp dụng ở nhiều lĩnh vực khác nhau, bao gồm nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên, dự đoán thị trường tài chính, điều khiển xe tự hành, và nhiều lĩnh vực khác nữa.

Machine Learning sử dụng đa dạng các thuật toán và phương pháp phân tích dữ liệu như regressions (hồi quy), decision trees (cây quyết định), neural networks (mạng nơ-ron), support vector machines và clustering. Các công cụ và thư viện phổ biến trong Machine Learning bao gồm TensorFlow, Scikit-learn, Keras và PyTorch.

2.2.1. Tổng quan về Supervised Learning (Học có giám sát)

Học có giám sát (Supervised learning) là một phương pháp trong lĩnh vực máy học, trong đó mô hình học cách nhận biết các mẫu và quan hệ giữa các biến đầu vào (input) và biến đầu ra (output) từ dữ liệu huấn luyện đã được gán nhãn. Mục tiêu của học có giám sát là huấn luyện mô hình để nắm bắt các mẫu, quy luật, và thông tin hữu ích từ các cặp dữ liệu input-output đã biết trước.

Trong học có giám sát, dữ liệu huấn luyện bao gồm các cặp input-output, trong đó input là thông tin đầu vào và output là nhãn hoặc giá trị dự đoán tương ứng. Mô hình học từ các ví dụ này bằng cách điều chỉnh các tham số nội bộ để giảm thiểu sai lệch giữa dự đoán của mô hình và nhãn thực tế.

Có hai loại chính của học có giám sát là phân loại (classification) và hồi quy (regression). Trong phân loại, mục tiêu là dự đoán các nhãn rời rạc và phân loại các mẫu vào các nhóm khác nhau. Trong hồi quy, mục tiêu là dự đoán các giá trị liên tục và xác định mối quan hệ số học giữa các biến.

Có nhiều thuật toán phổ biến được sử dụng trong học có giám sát, bao gồm cây quyết định (decision trees), rừng ngẫu nhiên (random forests), máy vector hỗ trợ (support vector machines - SVM), naive Bayes, hồi quy logistic (logistic regressio

2.2.2. Tổng quan về Unsupervised Learning (Học không giám sát)

Học không giám sát (Unsupervised Learning) là một dạng của học máy trong đó các thuật toán học cách tạo ra mô hình từ dữ liệu không được gán nhãn. Khác với học giám sát (Supervised Learning), ở đó chúng ta cung cấp dữ liệu được gán nhãn, trong học không giám sát, các thuật toán phải tự tìm ra các mẫu và mối quan hệ từ dữ liệu không có nhãn.

Mục tiêu của học không giám sát là tìm ra các mẫu ẩn, nhóm các điểm dữ liệu tương tự hoặc giảm số chiều của dữ liệu. Điều này có thể hữu ích trong việc phân cụm, phát hiện bất thường, trực quan hóa dữ liệu và học đặc trưng.

Có một số thuật toán phổ biến được sử dụng trong học không giám sát, bao gồm:

* Thuật toán gom cụm (clustering), như k-means, hierarchical clustering và DBSCAN, nhóm các điểm dữ liệu tương tự với nhau dựa trên độ tương đồng.
* Kỹ thuật giảm số chiều (dimensionality reduction), như phân tích thành phần chính (PCA) và t-SNE, chuyển đổi dữ liệu từ không gian có số chiều cao xuống một không gian có số chiều thấp hơn trong khi vẫn giữ lại cấu trúc quan trọng của dữ liệu.
* Thuật toán học luật kết hợp (association rule learning), như Apriori và FP-growth, phát hiện các mối quan hệ và liên kết thú vị giữa các biến trong tập dữ liệu.

Học không giám sát được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm tài chính, marketing, chăm sóc sức khỏe và phân tích mạng xã hội. Nó cho phép chúng ta tìm ra thông tin từ lượng lớn dữ liệu không được gán nhãn và khám phá các mẫu ẩn mà ban đầu có thể không dễ nhìn thấy.

2.2.3. Tổng quan về Reinforcement Learning (Học tăng cường)

Học tăng cường (Reinforcement Learning) là một loại thuật toán máy học cho phép tác nhân học cách đưa ra quyết định trong một môi trường để đạt được một mục tiêu cụ thể. Không giống như học có giám sát, trong đó mô hình được đào tạo dựa trên các ví dụ được gắn nhãn, Reinforcement Learning học thông qua thử và sai bằng cách tương tác với môi trường.

Trong Reinforcement Learning, một tác nhân thực hiện các hành động trong một môi trường và nhận phản hồi dưới dạng phần thưởng hoặc hình phạt, điều này cho biết chất lượng của các hành động đó. Mục tiêu của tác nhân là tìm hiểu một chính sách, là ánh xạ từ trạng thái đến hành động, nhằm tối đa hóa phần thưởng tích lũy theo thời gian.

Quá trình Reinforcement Learning thường bao gồm các thành phần sau:

* Environment (Môi trường): Hệ thống bên ngoài hoặc mô phỏng trong đó tác nhân hoạt động.
* State (Trạng thái): Điều kiện hiện tại hoặc đại diện của môi trường.
* Action (Hành động): Quyết định hoặc lựa chọn được thực hiện bởi tác nhân dựa trên trạng thái hiện tại.
* Reward (Phần thưởng): Tín hiệu phản hồi do môi trường cung cấp để đánh giá hành động của tác nhân.
* Policy (Chính sách): Chiến lược hoặc quy tắc mà tác nhân tuân theo để chọn hành động dựa trên trạng thái.
* Value function (Hàm giá trị): Ước tính phần thưởng tích lũy dự kiến ​​cho một trạng thái nhất định hoặc cặp trạng thái-hành động.
* Q-function (Chức năng Q): Một chức năng ước tính phần thưởng tích lũy dự kiến ​​cho một cặp hành động trạng thái nhất định.

Có nhiều thuật toán khác nhau được sử dụng trong Reinforcement Learning, chẳng hạn như Q-Learning, SARSA và Deep Q-Networks (DQN), mỗi thuật toán đều có điểm mạnh và hạn chế riêng. Reinforcement Learning đã được áp dụng thành công cho nhiều lĩnh vực khác nhau, bao gồm chế tạo người máy, chơi trò chơi và lái xe tự hành.

Nhìn chung, Reinforcement Learning cung cấp một khuôn khổ mạnh mẽ để cho phép các tổng đài viên tìm hiểu các chiến lược ra quyết định tối ưu trong môi trường phức tạp và năng động bằng cách tương tác với môi trường và nhận phản hồi dưới dạng phần thưởng hoặc hình phạt.

2.3. Tổng quan về Regression trong Machine Learning

2.3.1. Linear Regression (Hồi Quy Tuyến Tính)

Linear Regression (Hồi Quy Tuyến Tính) là một phương pháp trong Machine Learning và thống kê được sử dụng để mô hình hóa mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc. Đây là một trong những phương pháp đơn giản nhất và phổ biến nhất trong học máy.

Các Đặc Điểm Cơ Bản của Linear Regression:

* Biến Phụ Thuộc (Dependent Variable): Là biến mà chúng ta muốn dự đoán. Trong Linear Regression, biến này thường là liên tục, ví dụ như giá cả, doanh số bán hàng, v.v.
* Biến Độc Lập (Independent Variable): Là các biến được sử dụng để dự đoán biến phụ thuộc. Chúng có thể là một hoặc nhiều biến, và thường là dữ liệu số.
* Mô Hình Hồi Quy Tuyến Tính: Mô hình dựa trên giả định rằng mối quan hệ giữa các biến độc lập và biến phụ thuộc là tuyến tính. Trong trường hợp đơn giản nhất với một biến độc lập, mô hình có dạng:

*y = mx + b*

Trong đó, *y* là biến phụ thuộc, *x* là biến độc lập, *m* là độ dốc của đường thẳng (slope), và *b* là điểm cắt trục tung (intercept).

Quá Trình Huấn Luyện Mô Hình:

* Tìm Đường Thẳng Tốt Nhất: Mục tiêu của huấn luyện là tìm ra đường thẳng (hoặc siêu mặt phẳng trong trường hợp đa chiều) sao cho sai số dự đoán của mô hình là nhỏ nhất.
* Phương Pháp Tối Ưu Hóa: Thông thường, Linear Regression sử dụng phương pháp tối ưu hóa như Ordinary Least Squares (OLS) để tìm ra các giá trị \(m\) và \(b\) tối ưu nhất.

Đánh Giá Mô Hình:

* Mean Squared Error (MSE) và RMSE: Dùng để đánh giá hiệu suất của mô hình. MSE đo lường trung bình của bình phương sai số giữa dự đoán và giá trị thực tế, trong khi RMSE là căn bậc hai của MSE.
* R-squared (R²): Đo lường mức độ biến thiên của biến phụ thuộc được giải thích bởi mô hình. Giá trị R² càng gần 1 thể hiện mô hình giải thích tốt dữ liệu.

2.3.2. Random Forest Regression (Rừng Ngẫu Nhiên Hồi Quy)

Random Forest Regression là một phương pháp học máy trong đó một loạt các cây quyết định (decision trees) được tạo ra và kết hợp lại để tạo thành một mô hình dự đoán. Khác với việc sử dụng một cây quyết định duy nhất, Random Forest sử dụng nhiều cây và kết hợp kết quả từ các cây này để đưa ra dự đoán chính xác hơn.

Đặc Điểm Cơ Bản của Random Forest Regression:

* Tạo Ra Nhiều Cây Quyết Định (Decision Trees): Random Forest tạo ra một tập hợp các cây quyết định độc lập nhau, mỗi cây được huấn luyện trên một tập dữ liệu con và một số lượng nhỏ các đặc trưng được chọn ngẫu nhiên.
* Phương Pháp Bầu Cử (Voting Method): Khi có dữ liệu mới cần dự đoán, mỗi cây trong Random Forest sẽ đưa ra dự đoán của mình, và kết quả cuối cùng sẽ dựa trên việc "bầu cử" từ tất cả các cây để đưa ra kết quả cuối cùng.
* Ứng Dụng Trong Hồi Quy: Trong trường hợp của Regression, Random Forest Regression tính trung bình kết quả của tất cả các cây để đưa ra dự đoán.

Điểm Mạnh và Ứng Dụng:

* Khả Năng Xử Lý Dữ Liệu Lớn: Random Forest có khả năng xử lý các tập dữ liệu lớn và có nhiều đặc trưng mà không cần nhiều tiền xử lý dữ liệu.
* Chống Overfitting (Quá Mức Khớp Dữ Liệu): Bởi vì kết quả cuối cùng được tính từ nhiều cây khác nhau, nó thường ít bị overfitting hơn so với một cây quyết định duy nhất.
* Ứng Dụng Rộng Rãi: Random Forest có thể được áp dụng trong nhiều bài toán hồi quy như dự đoán giá nhà, dự đoán doanh thu, hoặc các bài toán có dữ liệu số lượng lớn.

Random Forest Regression là một phương pháp mạnh mẽ và linh hoạt trong việc dự đoán giá trị liên tục dựa trên dữ liệu đa chiều. Sự kết hợp của nhiều cây quyết định giúp cải thiện độ chính xác và ổn định của mô hình.

2.4. Support Vector Machine (SVM)

Support Vector Machine (SVM) là một thuật toán học máy có khả năng giải quyết cả bài toán phân lớp và hồi quy. Đặc biệt, nó nổi tiếng với khả năng xử lý tốt các bộ dữ liệu lớn, có tính phi tuyến tính cao và có khả năng xử lý số chiều dữ liệu cao.

2.4.1. Cơ Bản về Support Vector Machine

Mục Tiêu Chính: Support Vector Machine (SVM) là tìm ra siêu phẳng (hyperplane) tốt nhất để phân chia các lớp dữ liệu trong không gian nhiều chiều. Đây là siêu phẳng có khả năng tối đa hóa margin, tức là khoảng cách từ siêu phẳng đến các điểm dữ liệu gần nhất của các lớp là lớn nhất.

Mục tiêu này thể hiện sự cố gắng của SVM để tạo ra một ranh giới phân chia giữa các lớp dữ liệu một cách tối ưu, tức là một ranh giới mà có thể tối đa hoá khoảng cách giữa các điểm dữ liệu gần nhất của các lớp, điều này giúp tăng tính tổng quát và hiệu suất của mô hình. Điểm gần nhất từ các lớp đến siêu phẳng được gọi là các "support vectors" và chúng định rõ vị trí và hình dạng của siêu phẳng.

Phân Loại Phi Tuyến Tính: SVM có thể áp dụng các hàm nhân (kernel functions) để biến đổi không gian dữ liệu ban đầu thành không gian có thể tuyến tính hoặc phi tuyến tính, giúp nó phân loại dữ liệu không thể phân chia tuyến tính trong không gian ban đầu.

2.4.2. Các Khái Niệm Quan Trọng trong SVM

Siêu Phẳng (Hyperplane): Là một siêu biến (hyperplane) trong không gian nhiều chiều mà nó phân chia hai lớp dữ liệu. Trong không gian hai chiều, đó có thể là một đường thẳng, trong không gian ba chiều, đó là một mặt phẳng, và trong không gian nhiều chiều hơn, đó là một siêu phẳng.

Hàm Nhân (Kernel Function): Hàm nhân trong Machine Learning là một công cụ để biến đổi không gian dữ liệu ban đầu sang một không gian mới, giúp SVM và các phương pháp học máy khác phân loại dữ liệu phi tuyến tính. Các loại hàm nhân như Linear, Polynomial, và RBF Kernel được sử dụng để chuyển đổi không gian dữ liệu, tăng khả năng phân loại của các thuật toán trong học máy.Margin và Support Vectors: Margin là khoảng cách từ siêu phẳng tới điểm gần nhất của các lớp. Support Vectors là các điểm dữ liệu gần siêu phẳng và quyết định hình dạng và vị trí của siêu phẳng.

Regularization và Các Tham Số: Regularization trong Machine Learning là một kỹ thuật để kiểm soát và ngăn chặn overfitting trong quá trình huấn luyện mô hình. Nó thêm một thành phần vào hàm mất mát của mô hình để giảm thiểu sự phức tạp của mô hình, thông qua việc điều chỉnh trọng số và tránh việc chúng trở nên quá lớn hoặc không cần thiết, từ đó tạo ra một mô hình có khả năng tổng quát hoá tốt hơn cho dữ liệu mới.

2.4.3. Ứng Dụng và Ưu Điểm

Phân Lớp và Hồi Quy: SVM có thể áp dụng vào cả bài toán phân loại (classification) và hồi quy (regression).

Phân lớp và hồi quy là hai kỹ thuật quan trọng trong Machine Learning được áp dụng rộng rãi trong nhiều lĩnh vực khác nhau với các ưu điểm và ứng dụng riêng.

Phân Lớp:

* Ứng Dụng: Dự Đoán và Phân Loại: Phân lớp được sử dụng để phân loại dữ liệu vào các nhóm hoặc lớp khác nhau. Ví dụ, trong phát hiện email rác, dự đoán bệnh lý dựa trên triệu chứng, nhận dạng hình ảnh và nhiều ứng dụng khác.
* Ưu Điểm:
* Dễ Hiểu và Áp Dụng: Có thể áp dụng các thuật toán đơn giản như K-nearest neighbors (KNN) hoặc các mô hình phức tạp như Support Vector Machines (SVM) cho nhiều loại dữ liệu.
* Hiệu Suất Cao: Các thuật toán phân lớp có thể tạo ra mô hình có độ chính xác cao khi có dữ liệu đủ mạnh và cân nhắc đúng tham số.

Hồi Quy:

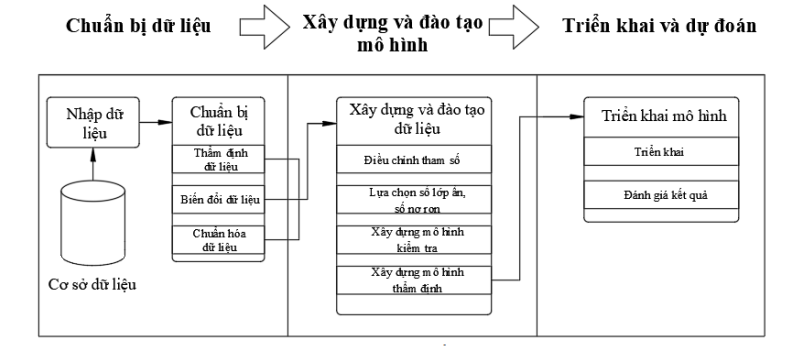
* Ứng Dụng:
* Dự Đoán Giá Trị Liên Tục Hồi quy được sử dụng để dự đoán các giá trị liên tục như giá cổ phiếu, giá nhà, doanh số bán hàng dựa trên các biến đầu vào.
* Ưu Điểm:
* Dự Đoán Chính Xác: Cho phép dự đoán giá trị cụ thể dựa trên một số đặc trưng đã biết.
* Phân Tích Liên Kết: Hồi quy có thể giúp hiểu mối quan hệ giữa các biến đầu vào và đầu ra.

Cả hai kỹ thuật này đều là công cụ mạnh mẽ trong Machine Learning, giúp phân tích, dự đoán và hiểu sâu hơn về dữ liệu trong nhiều ngữ cảnh khác nhau. Sự lựa chọn giữa phân lớp và hồi quy phụ thuộc vào loại dữ liệu và vấn đề cụ thể mà bạn đang xử lý.

Hiệu Suất Cao Trên Dữ Liệu Lớn: SVM có thể xử lý hiệu quả các tập dữ liệu lớn, có khả năng tốt khi có số lượng lớn các đặc trưng và dữ liệu nhiễu.

CHƯƠNG 3. MÔ HÌNH THỰC NGHIỆM

3.1. Quy trình thực hiện mô hình



Hình 3. 1: Quy trình thực hiện mô hình

Quy trình thực hiện một mô hình học máy thường gồm các bước:

Thu thập và Khám phá Dữ liệu:

Bước đầu tiên là thu thập dữ liệu liên quan đến bài toán. Dữ liệu có thể đến từ nhiều nguồn khác nhau, bao gồm cơ sở dữ liệu, tệp văn bản, dữ liệu hình ảnh, hoặc dữ liệu trực tuyến.

Sau khi thu thập, dữ liệu cần được khám phá để hiểu rõ về tính chất của nó. Điều này bao gồm kiểm tra dữ liệu bị thiếu, phân phối của biến, và tương quan giữa các biến. Các biểu đồ và số liệu thống kê có thể được sử dụng để trực quan hóa thông tin.

Tiền Xử Lý Dữ liệu:

Dữ liệu thường cần phải được tiền xử lý trước khi đưa vào mô hình. Điều này bao gồm xử lý dữ liệu bị thiếu bằng cách loại bỏ hoặc điền giá trị thiếu, chuẩn hóa dữ liệu để đảm bảo mọi biến có cùng phạm vi giá trị, và mã hóa biến phân loại thành dạng số.

Chia Dữ liệu:

Dữ liệu cần được chia thành tập huấn luyện và tập kiểm tra. Tập huấn luyện được sử dụng để huấn luyện mô hình, trong khi tập kiểm tra được sử dụng để đánh giá hiệu suất của mô hình. Kỹ thuật phân chia có thể là giữ lại tỷ lệ cố định (ví dụ: 70% huấn luyện, 30% kiểm tra) hoặc cross-validation.

Lựa Chọn Mô Hình:

Lựa chọn mô hình là bước quan trọng. Điều này phụ thuộc vào loại dữ liệu và bài toán cụ thể. Mô hình có thể là hồi quy tuyến tính, cây quyết định, mạng nơ-ron, hay mô hình học sâu khác như CNN hoặc RNN.

Huấn Luyện Mô Hình:

Mô hình được huấn luyện trên tập dữ liệu huấn luyện. Quá trình này bao gồm tối ưu hóa các tham số của mô hình để nó có thể dự đoán hoặc phân loại dữ liệu một cách chính xác.

Đánh Giá Mô Hình:

Mô hình cần được đánh giá bằng cách sử dụng tập kiểm tra hoặc cross-validation. Các phép đo đánh giá hiệu suất bao gồm accuracy, precision, recall, F1-score, và ROC-AUC tùy thuộc vào loại bài toán.

Tinh Chỉnh Mô Hình:

Nếu mô hình không đạt hiệu suất mong muốn, bạn có thể cần phải tinh chỉnh các siêu tham số của mô hình hoặc thay đổi kiến trúc mô hình.

Triển Khai Mô Hình:

Khi mô hình đạt hiệu suất tốt và được chấp nhận, nó có thể được triển khai trong môi trường thực tế để sử dụng trong các ứng dụng hàng ngày.

Theo Dõi và Cải Thiện:

Mô hình cần được theo dõi trong môi trường sản xuất và cải thiện theo thời gian khi có thêm dữ liệu mới và khi cần thiết.

3.2. Các kỹ thuật chuẩn bị dữ liệu sử dụng

**SimpleImputer** là một lớp trong thư viện scikit-learn của python, được sử dụng để xử lý các giá trị thiếu (missing values) trong dữ liệu. SimpleImputer thực hiện các chiến lược thay thế giá trị thiếu bằng một giá trị cụ thể hoặc thông qua các phương pháp thống kê đơn giản như trung bình , trung vị , hoặc giá trị xuất hiện nhiều nhất.

**OneHotEncoder** dùng để chuyển đổi biến phân loại thành biểu diễn số học

**OrdinalEncoder** dùng để chuyển đổi biến phân loại thành biểu diễn số học

**SelectKBest** dùng để chọn các đặc trưng tốt nhất từ một tập dữ liệu dựa trên các phép đo thống kê, nó được sử dụng trong quá trình chọn đặc trưng để cải thiện hiệu suất của mô hình và giảm chiều dữ liệu

**StandardScaler** là kỹ thuật chuẩn hóa dữ liệu, được sử dụng để chuyển đổi các biến đặc trưng sao cho chúng có độ lệch chuẩn bằng 0 và giá trị trung bình bằng 0. Điểu này làm mềm dữ liệu và đảm bảo rằng tất cả các đặc trưng có cùng quy mô, giúp mô hình học máy hội tụ nhanh hơn và dễ dàng tìm kiếm các mối quan hệ trong dữ liệu

3.3. Tông quan bộ dữ liệu

Bộ dữ liệu chứa 13 đặc trưng

Id: Để đếm các bản ghi

MSSubClass: Xác định loại nhà ở liên quan đến việc bán

MSZoning: Xác định phân hoại quy hoạch chung của việc bán hàng

LotArea: Kích thước lô tính bằng feet

LotConfig: Cấu hình của lô

BldgType: Loại nhà ở

OverallCond: Đánh giá tình trạng chung của ngôi nhà

YearBuilt: Năm xây dựng ban đầu

YearRemodAdd: Ngày tu sửa (bằng ngày xây dựng nếu không tu sửa hoặc bổ sung)

Exterior1st: Lớp phủ bên ngoài ngôi nhà

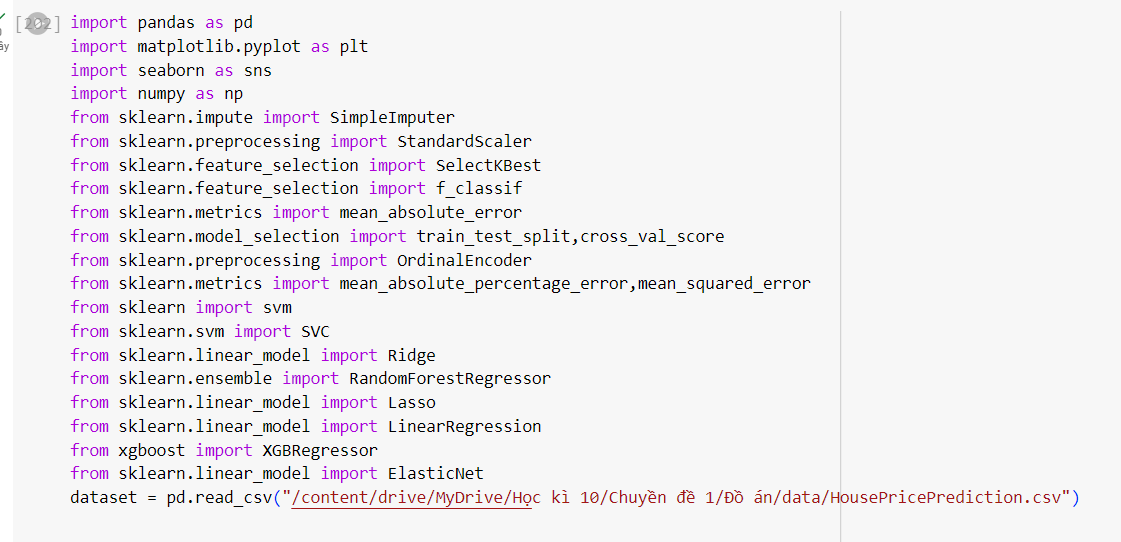
BsmtFinSF2: Diện tích sàn tầng hầm đã hoàn thành

TotalBsmtSF: Tổng diện tích sàn tầng hầm

SalePrice: Giá của căn nhà

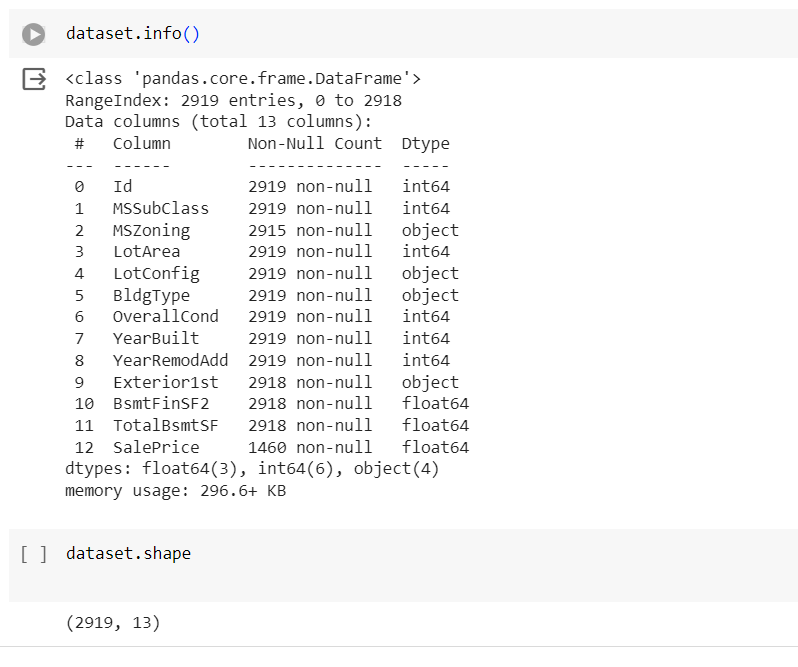
3.4. Thực nghiệm mô hình

Nhập các thư viện và tập dữ liệu



Hình 3. 2: Nhập các thư viện và tập dữ liệu

Thông tin và kích thước của tập dữ liệu



Hình 3. 3: Thông tin và kích thước của tập dữ liệu

**Tiền xử lý dữ liệu**

Bây giờ, chúng ta phân loại các đặc trưng tùy thuộc vào kiểu dữ liệu của chúng (int,float,object) và sau đó tính toán số lượng của chúng

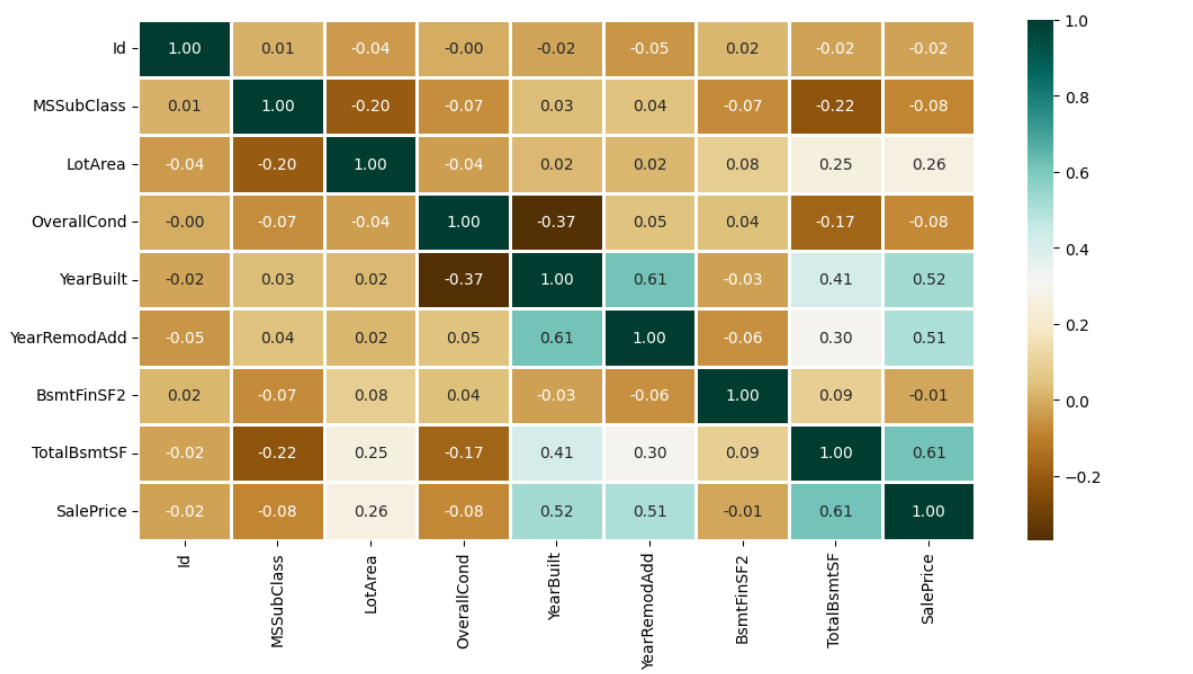


Hình 3. 4: Ví dụ code phân loại các đặc trưng

**Exploratory Data Analysis**

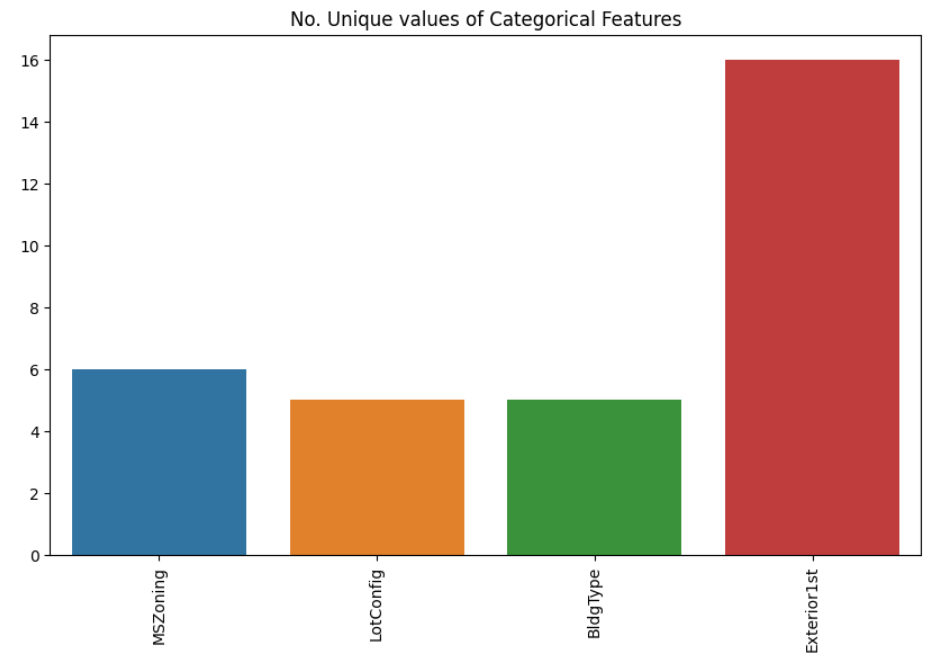
EDA đề cập đến việc phân tích sâu dữ liệu để khám phá các mẫu khác nhau và phát hiện các điểm bất thường. Trước khi đưa ra suy luận từ dữ liệu, điều cần thiết là phải kiểm tra tất cả các biến

Vì vậy ở đây chúng ta tạo một bản đồ nhiệt bằng thư viện seaborn



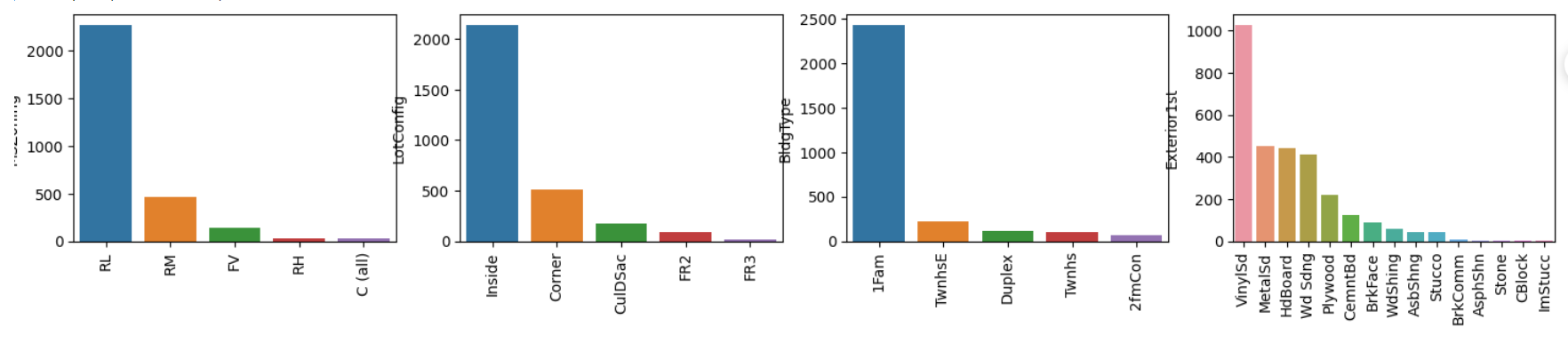
Hình 3. 5: Tương quan của các đặc trưng

Để phân tích các đặc trưng phân loại khác nhau. Vẽ barplot



Hình 3. 6: Các giá trị unique của các đặc trưng Categorical

Cho thấy Exterior1st có khoảng 16 danh mục duy nhất và các đặc trưng khác có khoảng 6 danh mục duy nhất. Để tìm ra số lượng thực tế của từng danh mục, chúng ta có thể vẽ biểu đồ của từng đặc điểm riêng biệt

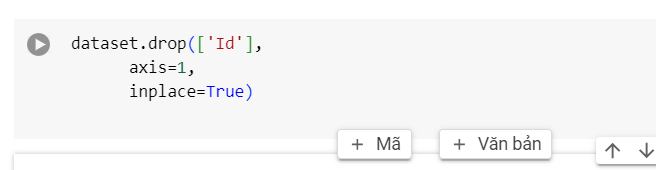


Hình 3. 7: Số lượng thực tế của các đặc trưng Categorical

**Làm sạch dữ liệu**

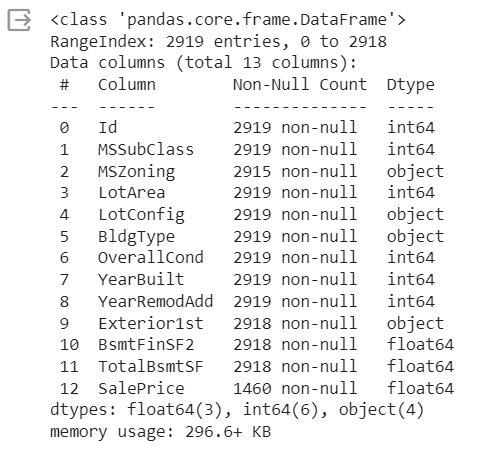
Làm sạch dữ liệu là cách ứng biến dữ liệu hoặc loại bỏ dữ liệu không chính xác, bị hỏng hoặc không liên quan

Như trong tập dữ liệu của chúng tôi, có một số cột không quan trọng và không liên quan đến việc đào tạo mô hình. Vì vậy chúng ta có thể bỏ cột đó trước khi đào tạo

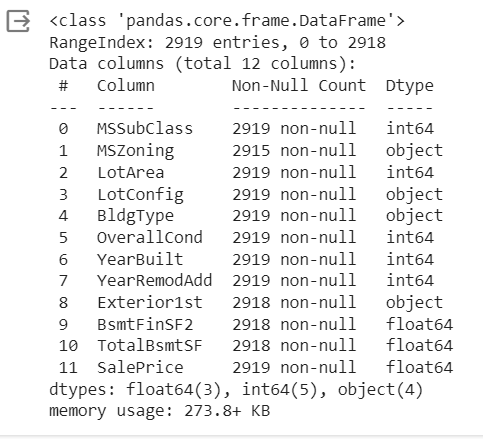


Hình 3. 8: Ví dụ code làm sạch dữ liệu

Sử dụng lớp SimpleImputer để thay thế các giá trị trống của SalePrice bằng giá trị trung bình của chúng để làm cho công việc phân phối dữ liệu trở nên đối xứng



Hình 3. 9: Dữ liệu trước khi xử lý



Hình 3. 10: Dữ liệu sau khi xử lý

Có thế thấy các dữ liệu trống của SalePrice đã được thay thế bằng giá trị trung bình

Vẫn còn một số đặc trưng có giá trị trống, ta thực hiện xóa các dòng bị thiếu và kiểm tra dữ liệu còn thiếu không



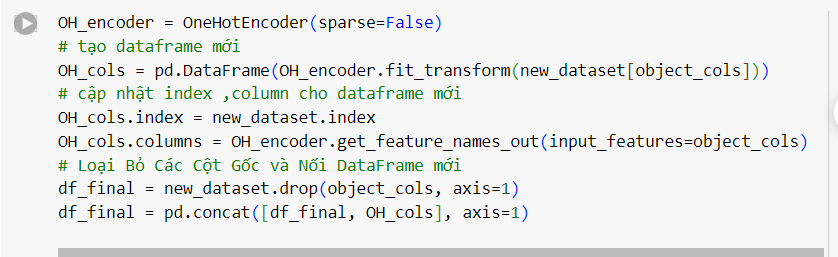
Hình 3. 11: Ví dụ code xóa các dòng bị thiếu và kiểm tra dữ liệu còn thiếu không

Sử dụng OneHotEncode dành cho các đặc trưng phân loại



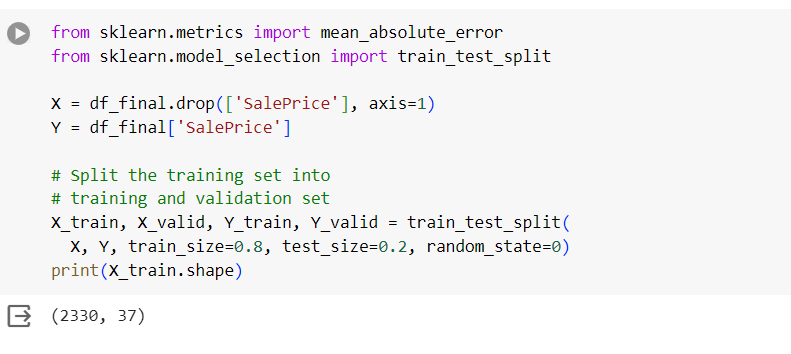
Hình 3. 12: Ví dụ code sử dụng OneHotEncode

Sau đó, khi chúng ta có danh sách tất cả các đặc trưng. Chúng ta có thể áp dụng OneHotEncoding cho toàn bộ danh sách



Hình 3. 13: OneHotEncoding cho toàn bộ danh sách

Chia dữ liệu thành tập train và test



Hình 3. 14: Ví dụ code chia tập dữ liệu thành train và test

**Mô hình và độ chính xác**

**\*Mô hình**

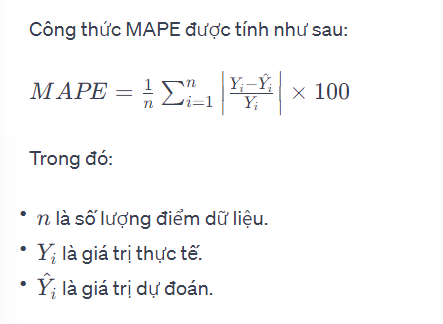
Vì chúng ta phải huấn luyện mô hình để xác định các giá trị liên tục nên chúng ta sẽ sử dụng các mô hình hồi quy như

* SVM: Support Vector Machine là một thuật toán học máy sử dụng để phân loại hoặc dự đoán giá trị. Nó tìm một ranh giới quyết định giữa các lớp dữ liệu, cố gắng tối đa hóa khoảng cách giữa các điểm dữ liệu gần nhất và ranh giới
* Ridge: Thuật toán Ridge là một phương pháp hồi quy tuyến tính sử dụng để giảm overfitting trong mô hình hồi quy. Cố gắng giữ các trọng số của biến độc lập càng nhỏ có thể, Ridge thêm một thành phần độ lớn vào hàm mất mát, được kiểm soát bởi tham số alpha
* Random Forest Regressor: RandomForestRegressor là một mô hình học máy sử dụng nhiều cây quyết định để dự đoán giá trị liên tục. Nó giảm overfitting bằng cách xây dựng các cây độc lập trên tập con dữ liệu ngẫu nhiên và kết hợp dự đoán của chúng
* Lasso: Thuật toán Lasso là một phương pháp hồi quy tuyến tính được sử dụng để giảm overfitting và chọn lọc biến độc lập. Nó thêm một thành phần độ lớn vào hàm mất mát, kiểm soát bởi tham số alpha, và có khả năng đặt các trọng số của một số biến độc lập bằng không, từ đó giúp "chọn lọc" các biến quan trọng
* Linear Regression: là một mô hình học máy dùng để dự đoán giá trị dựa trên mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc. Mục tiêu là tìm một đường thẳng (phương trình tuyến tính) sao cho tổng bình phương sai số giữa giá trị dự đoán và thực tế là nhỏ nhất
* XGBRegressor là một mô hình học máy thuộc dạng Gradient Boosting, chủ yếu sử dụng trong bài toán dự đoán giá trị liên tục. Nó kết hợp nhiều cây quyết định để tạo ra một mô hình mạnh mẽ, sử dụng cơ chế cải tiến đặc biệt trong quá trình đào tạo để giảm thiểu hàm mất mát
* ElasticNet là một phương pháp hồi quy kết hợp cả Lasso và Ridge bằng cách thêm cả hai thành phần độ lớn vào hàm mất mát. Nó giúp kiểm soát overfitting, giữ lại các biến quan trọng, và đồng thời giải quyết vấn đề của Lasso khi có sự tương quan cao giữa các biến độc lập

**\*Các Metric đánh giá độ chính xác**

MAPE là viết tắt của "Mean Absolute Percentage Error" (Sai số Tuyệt đối Trung bình theo Tỷ lệ phần trăm). Đây là một phương pháp đo lường độ lệch tuyệt đối trung bình giữa giá trị dự đoán và giá trị thực tế, được thể hiện dưới dạng phần trăm của giá trị thực tế.

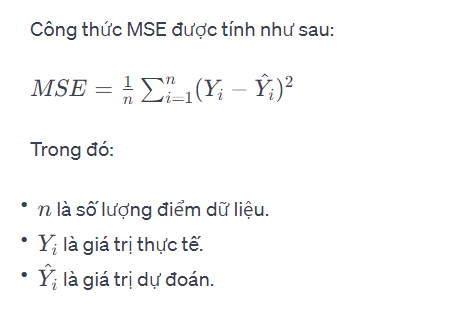
Công thức tính MAPE cho một tập dữ liệu gồm n điểm dữ liệu là



Hình 3. 15: Công thức tính MAPE

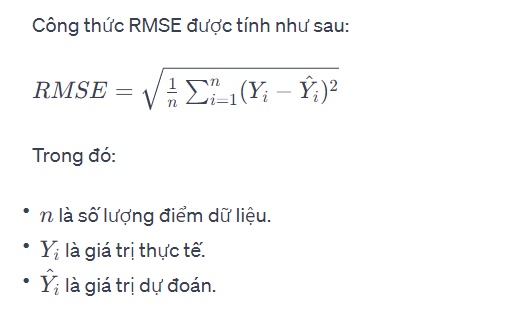
Mục tiêu là giữa cho giá trị MAPE thấp, vì giá trị càng thấp càng chỉ ra rằng mô hình dự đoán gần với giá trị thực tế. Giá trị MAPE 0% đồng nghĩa với việc mô hình không có lỗi dự đoán

MSE là viết tắt của "Mean Squared Error" (Sai số Bình phương Trung bình). Đây là một phương pháp đo lường độ lệch bình phương trung bình giữa giá trị dự đoán và giá trị thực tế.

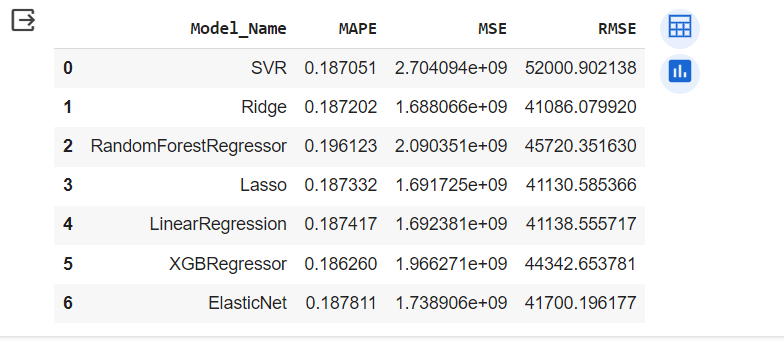


Hình 3. 16: Công thức MSE

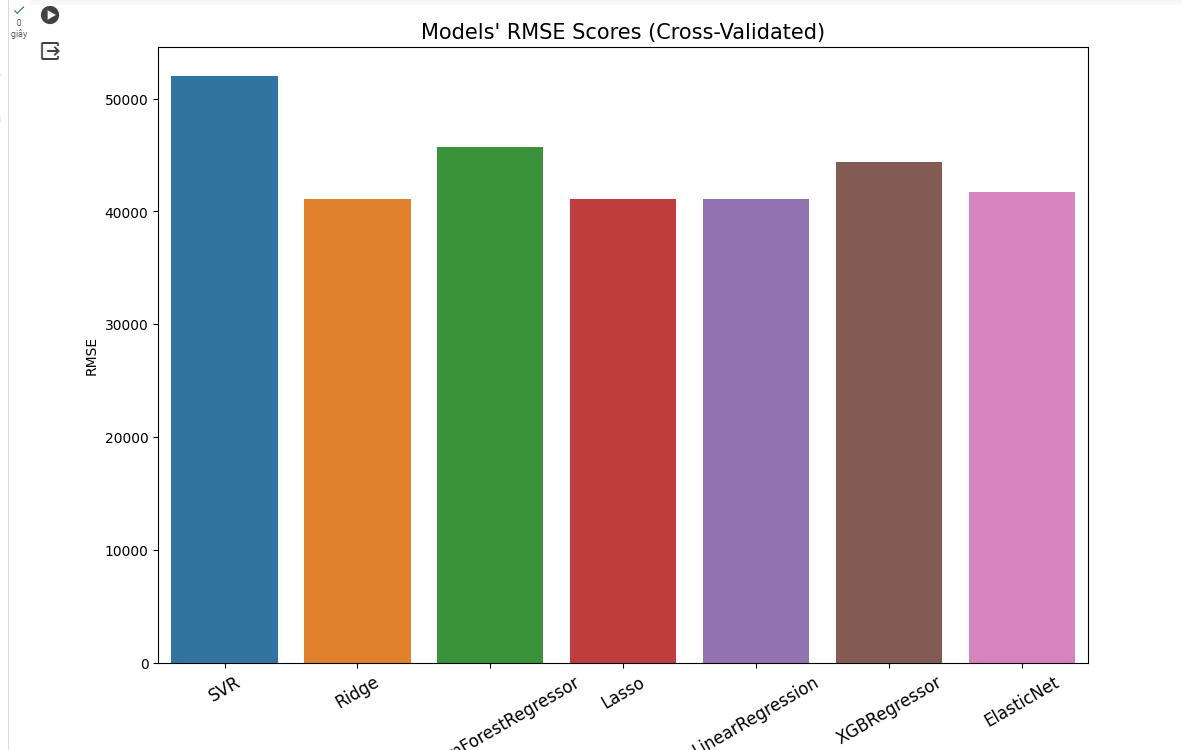
RMSE là viết tắt của "Root Mean Squared Error" (Sai số Bình phương Trung bình căn bậc hai). Đây là một phương pháp đo lường độ lệch trung bình giữa giá trị dự đoán và giá trị thực tế, với sự ưu tiên đối với các sai số lớn.



Hình 3. 17: Công thức MRSE

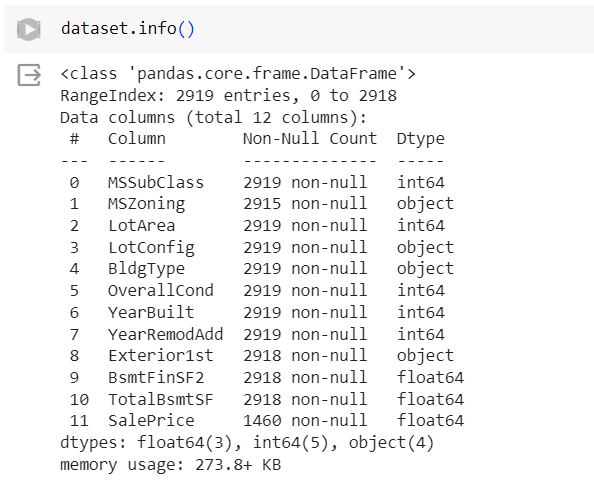


Hình 3. 18: Chỉ số đánh giá các mô hình



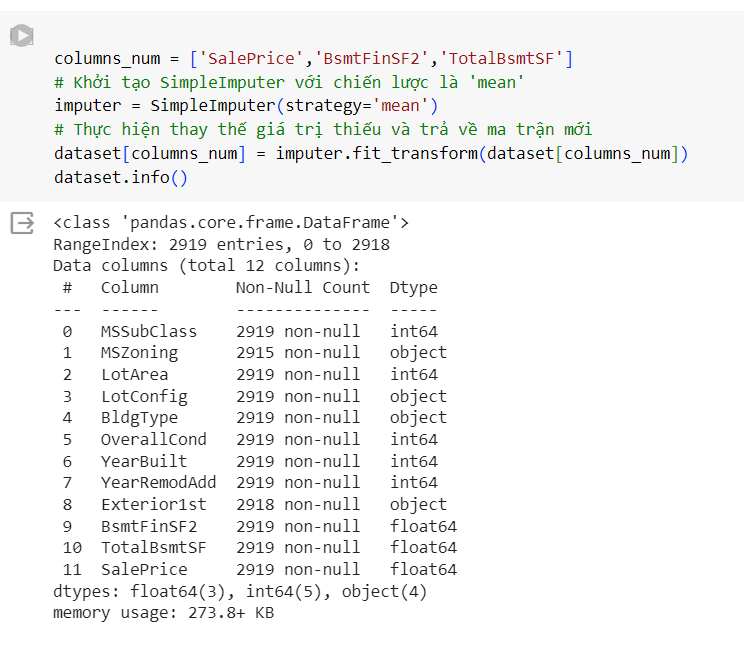
Hình 3. 19: RMSE Score của từng mô hình

**Sau đây ta có thể thay thế hoặc thêm các bước chuẩn bị dữ liệu để so hình với mô hình dự đoán trên nhằm tìm ra các kỹ thuật hay phương pháp để tối ưu bài toán**



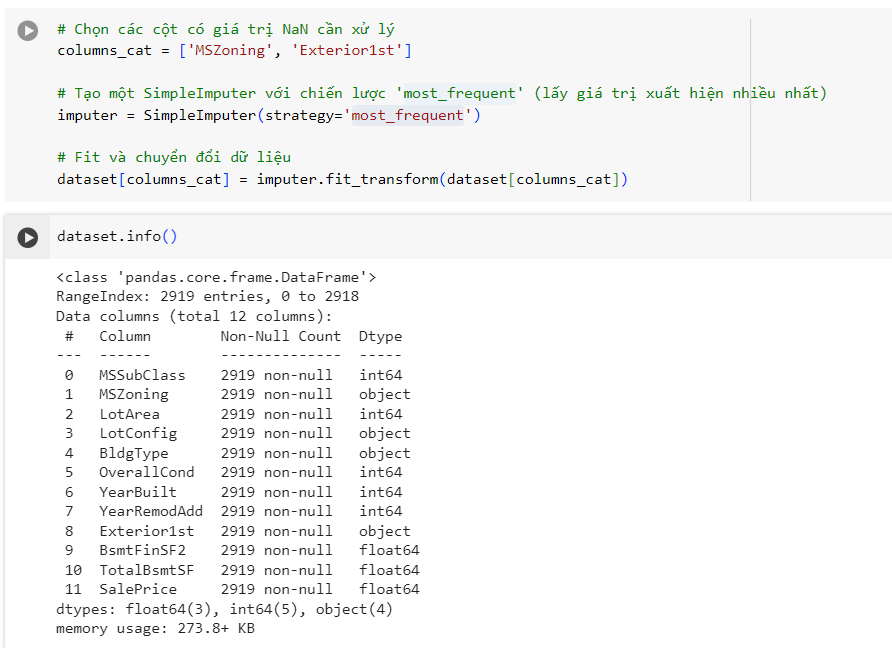
*Dữ liệu ban đầu*

Ta có thể thay thế các đặc trưng dữ liệu số bằng chiến lược ‘mean’ trong lớp SimpleImputer



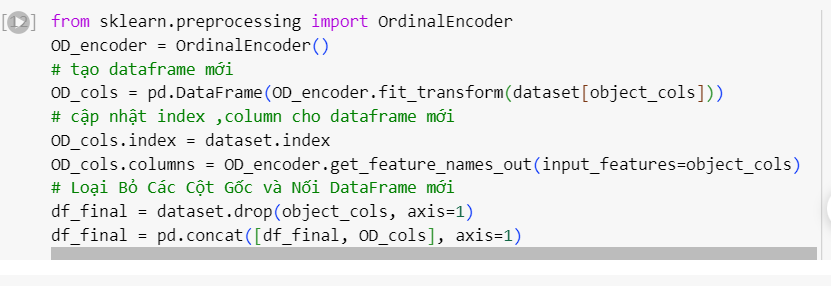
*Đặc trưng numerical được thay thế*

Và thay thế các giá trị thiếu trong các đặc trưng categorical bằng chiến lược ‘most\_frequent’



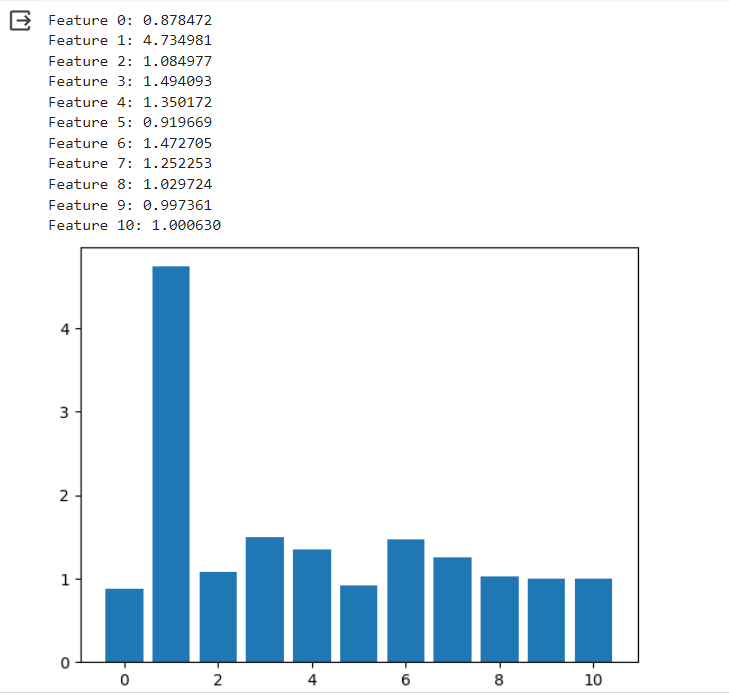
Đặc trưng categorical được thay thế

Ta có thể thay thế OneHotEncoder bằng OrdinalEncoder



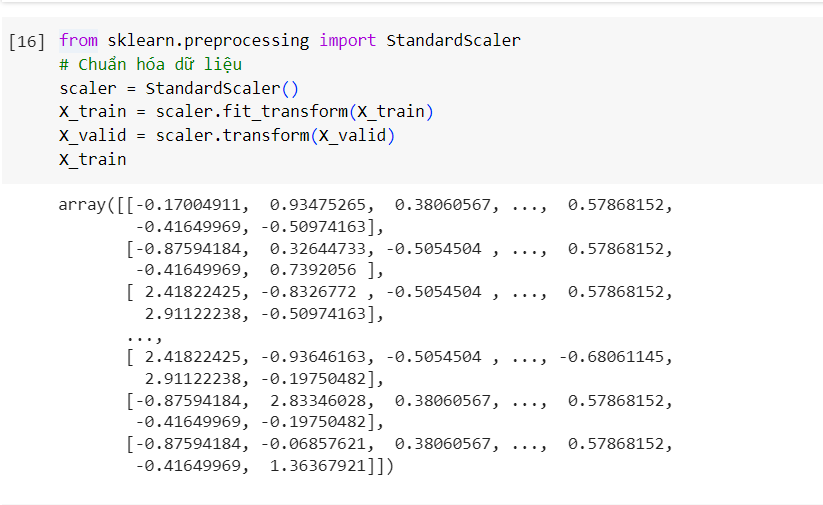
Hình 3. 20: thay thế OneHotEncoder bằng OrdinalEncoder

Sử dụng phương pháp SelectKBest để lựa chọn đặc trưng cho mô hình, ở đây có thể thấy độ quan trọng của các đặc trưng gần tương đương nhau nên ta chọn k =’all’



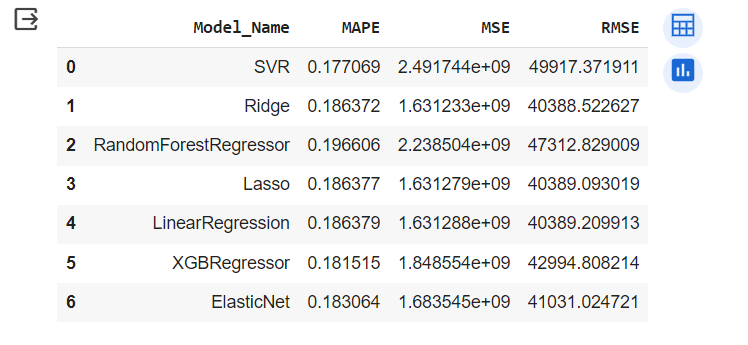
Hình 3. 21: Độ quan trọng của các đặc trưng

Sử dụng StandardScaler để chuẩn hóa dữ liệu

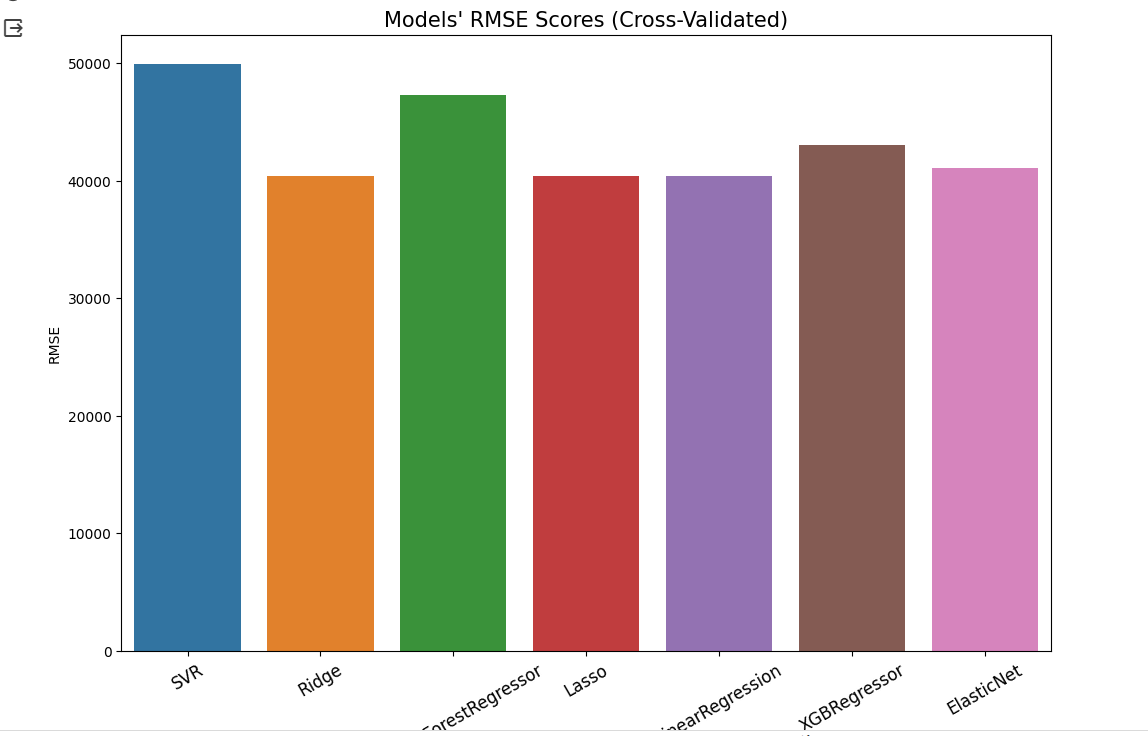


Hình 3. 22: Sử dụng StandardScaler để chuẩn hóa dữ liệu

Đánh giá mô hình



Hình 3. 23: Chỉ số đánh giá các mô hình sau khi thay đổi, thêm các kỹ thuật



Hình 3. 24: RMSE Score của từng mô hình sau khi thay đổi, thêm các kỹ thuật

3.5. Kết quả đạt được

Dựa trên các chỉ số đánh giá mô hình (MAPE, MSE, RMSE), ta có thể đưa ra một số nhận xét và kết luận chung về hiệu suất của các mô hình:

- Mô hình SVR có MAPE thấp nhất, cho thấy khả năng dự đoán tốt với tỷ lệ lỗi tương đối thấp.

- Cả Ridge Regression, Lasso Regression, và Linear Regression đều có hiệu suất tương tự, với MSE và RMSE ở mức thấp. Chúng thích hợp khi mô hình đơn giản và dễ giải thích là quan trọng.

- Random Forest và XGBoost Regressor có MSE và RMSE ổn định, nhưng MAPE cao hơn so với mô hình tuyến tính. Chúng có thể được xem xét khi cần độ chính xác cao hơn và có khả năng làm việc tốt với dữ liệu phi tuyến tính.

- ElasticNet, với sự kết hợp của L1 và L2 regularization, đưa ra kết quả trung bình với hiệu suất ổn định.

- Mô hình SVR đạt được MAPE thấp nhất nhưng với MSE và RMSE cao. Trong khi đó, các mô hình tuyến tính có MSE và RMSE thấp, nhưng MAPE cao hơn. Điều này thể hiện sự đánh đổi giữa độ chính xác tuyệt đối và giảm sai số bình phương trung bình.

CHƯƠNG 4. KẾT LUẬN

4.1. Kết luận

Trong quá trình phát triển và đánh giá các mô hình dự đoán giá nhà, tôi đã đạt được một loạt kết quả quan trọng. Các mô hình này đã chứng minh khả năng học và dự đoán giá nhà dựa trên dữ liệu lịch sử, và đã được đánh giá tích cực trên chỉ số MAPE, MSE và RMSE

Tuy nhiên, dự án vẫn còn rất nhiều tiềm năng để phát triển và nâng cấp. Các hướng phát triển tiềm năng bao gồm việc sử dụng dữ liệu bổ sung, thử nghiệm với các mô hình mới. Mặc dù mô hình dự đoán giá nhà là một công việc phức tạp và đầy thách thức, tôi hy vọng rằng nghiên cứu này có thể đóng góp vào sự hiểu biết về khả năng ứng dụng các mô hình hồi quy trong dự đoán giá nhà. Tôi kỳ vọng rằng việc phát triển và cải thiện mô hình sẽ giúp tạo ra một công cụ hữu ích cho các cá nhân, tổ chức khi chọn nhà

4.2. Hướng phát triển

Mặc dù mô hình đã đạt được kết quả tích cực, nhưng vẫn còn nhiều cơ hội để phát triển và nâng cao hiệu suất dự đoán. Dưới đây là một số hướng phát triển tiềm năng cho dự án trong tương lai:

Sử dụng Dữ liệu Bổ sung: Bổ sung dữ liệu cơ bản và kỹ thuật khác nhau

Sử dụng các kỹ thuật, thuật toán mới

Phát triển Ứng dụng Thực Tế: Tích hợp mô hình dự đoán giá nhà vào các ứng dụng thực tế cho các nhà đầu tư và người mua.

Đánh giá Rủi ro: Đối phó với tình huống không chắc chắn và rủi ro trong dự đoán giá nhà. Xác định các phương pháp quản lý rủi ro và áp dụng chúng trong mô hình.

Những hướng phát triển trên có thể giúp cải thiện tính chính xác và khả năng dự đoán của mô hình dự đoán giá nhà và làm cho nó trở thành một công cụ hữu ích cho các nhà đầu tư và người mua

TÀI LIỆU THAM KHẢO

[1]"Predicting House Prices Using Machine Learning: A Comprehensive Review"

* Tác giả: Nguyen, T., & Smith, J.
* Năm xuất bản: 2020
* Tên Tạp chí: International Journal of Machine Learning and Applications

[2]"A Comparative Analysis of Regression Models for Housing Price Prediction"

* Tác giả: Kim, H., & Patel, M.
* Năm xuất bản: 2019
* Tên Hội nghị: Proceedings of the Annual Conference on Artificial Intelligence

[3]"Enhancing Housing Price Prediction Models with Ensemble Learning"

* Tác giả: Lee, C., & Wang, S.
* Năm xuất bản: 2021
* Tên Tạp chí: Journal of Computational Intelligence in Real Estate

[4]"Incorporating Spatial Data in House Price Prediction Models"

* Tác giả: Gupta, R., & Chen, L.
* Năm xuất bản: 2018
* Tên Hội nghị: International Conference on Geospatial Data Science

[5]"Deep Learning Approaches for Real Estate Valuation"

* Tác giả: Wang, Y., & Li, Q.
* Tên Tạp chí: Neural Networks for Real Estate Research

[6]"Predictive Modeling of House Prices: A Practical Implementation Guide"

* Tác giả: Kim, J., & Nguyen, M.
* Năm xuất bản: 2017
* Tên Sách: Machine Learning in Real Estate

Link đồ án : <https://drive.google.com/drive/folders/13w3uAWai2IzpeqdB9TviuuK3kgVvvvFW?usp=sharing>