Big Data Course

Capstone Project   
Final Report

ⓒ2023 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

|  |
| --- |
| Phân Tích Dự Đoán Giá Nhà Đất |

**14/8/2024**

**Nhóm 1 – BD 02**

Lê Đông Anh Kiệt

Trần Tuấn Kiệt

Ngô Văn Lâu

Ngân Hoàng Huy

La Hồng Lộc

Nguyễn Đặng Phú Mẫn

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. Simulated Scenario Description

2.2. Datasets Selection and Description

2.3. Data Ingestion Pipeline

2.4. Data Transformation Processing

2.5. Data Query and Insight

3. Results

3.1. Data Ingestion Scripts and Code

3.2. Data Transformation Scripts and Code

3.3. Description and Sample of Transformed Datasets

3.4. Data Visualization of Query Results

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

1.1. Background Information

Trong những thập kỷ qua, sự phát triển vượt bậc của công nghệ thông tin đã mang lại nhiều thay đổi sâu rộng trong hầu hết các lĩnh vực kinh tế, và ngành bất động sản không phải là ngoại lệ. Trong bối cảnh thị trường bất động sản ngày càng trở nên phức tạp và cạnh tranh khốc liệt, việc dự đoán chính xác giá trị bất động sản trở thành một yếu tố quyết định đến sự thành công của các nhà đầu tư, nhà phát triển dự án, và các chuyên gia trong ngành. Trước đây, việc định giá bất động sản thường dựa vào kinh nghiệm cá nhân, cảm tính, và các phương pháp phân tích truyền thống, nhưng những phương pháp này thường không đủ để nắm bắt được tất cả các yếu tố tác động đến giá cả, từ tình trạng kinh tế vĩ mô đến các đặc điểm cụ thể của từng khu vực và từng tài sản.

Sự ra đời và phát triển của Big Data đã thay đổi hoàn toàn cách thức mà ngành bất động sản tiếp cận việc dự đoán giá nhà đất. Big Data cung cấp một lượng thông tin khổng lồ từ nhiều nguồn khác nhau, bao gồm dữ liệu về giao dịch bất động sản, dữ liệu kinh tế, xã hội, hạ tầng, và thậm chí cả dữ liệu cảm biến từ các thiết bị IoT. Thông qua việc xử lý và phân tích những khối dữ liệu lớn này, các thuật toán máy học (Machine Learning) có thể phát hiện ra những mối liên hệ phức tạp mà con người khó có thể nhận thấy, từ đó tạo ra các mô hình dự đoán với độ chính xác cao.

Không chỉ dừng lại ở việc cung cấp thông tin chi tiết về giá trị hiện tại của bất động sản, các mô hình dự đoán dựa trên Big Data còn có khả năng phân tích xu hướng và dự đoán các biến động giá trong tương lai. Điều này mang lại lợi ích to lớn cho các nhà đầu tư và nhà phát triển, giúp họ đưa ra các quyết định thông minh, tối ưu hóa lợi nhuận và giảm thiểu rủi ro trong các giao dịch. Ví dụ, các nhà đầu tư có thể sử dụng các công cụ dự đoán để xác định thời điểm mua vào hoặc bán ra phù hợp, lựa chọn các khu vực tiềm năng để đầu tư, hoặc đánh giá mức độ rủi ro của một dự án bất động sản cụ thể.

Tuy nhiên, việc áp dụng Big Data trong lĩnh vực bất động sản không phải là không gặp phải những thách thức. Một trong những thách thức lớn nhất là vấn đề bảo mật và quyền riêng tư dữ liệu. Với lượng dữ liệu khổng lồ được thu thập và xử lý, việc bảo vệ thông tin cá nhân và dữ liệu nhạy cảm của người dùng trở nên cấp thiết hơn bao giờ hết. Ngoài ra, độ chính xác của các mô hình dự đoán cũng phụ thuộc vào chất lượng của dữ liệu đầu vào, và việc thu thập, làm sạch và xử lý dữ liệu kh Mặc dù có những thách thức, không thể phủ nhận rằng sự kết hợp giữa công nghệ Big Data và ngành bất động sản đang mở ra những cơ hội chưa từng có. Những công nghệ này không chỉ giúp tối ưu hóa quy trình định giá mà còn hỗ trợ trong việc quản lý tài sản, phát triển dự án, và thậm chí là cá nhân hóa trải nghiệm người dùng. Trong tương lai, khi công nghệ tiếp tục phát triển, chúng ta có thể mong đợi rằng Big Data sẽ ngày càng đóng vai trò quan trọng hơn trong việc định hình thị trường bất động sản, đưa ra những dự đoán chính xác và tạo ra những giá trị mới cho tất cả các bên liên quan kông phải lúc nào cũng dễ dàng.

Mặc dù có những thách thức, không thể phủ nhận rằng sự kết hợp giữa công nghệ Big Data và ngành bất động sản đang mở ra những cơ hội chưa từng có. Những công nghệ này không chỉ giúp tối ưu hóa quy trình định giá mà còn hỗ trợ trong việc quản lý tài sản, phát triển dự án, và thậm chí là cá nhân hóa trải nghiệm người dùng. Trong tương lai, khi công nghệ tiếp tục phát triển, chúng ta có thể mong đợi rằng Big Data sẽ ngày càng đóng vai trò quan trọng hơn trong việc định hình thị trường bất động sản, đưa ra những dự đoán chính xác và tạo ra những giá trị mới cho tất cả các bên liên quan.

1.2. Motivation and Objective

Động lực chính của nhóm khi áp dụng Big Data vào việc dự đoán giá bất động sản là nhu cầu có một hiểu biết sâu sắc và chính xác hơn về thị trường bất động sản để đưa ra các quyết định đầu tư thông minh. Thị trường bất động sản luôn biến động và chịu ảnh hưởng bởi nhiều yếu tố phức tạp như kinh tế, chính trị, xã hội và môi trường. Sử dụng Big Data giúp thu thập và phân tích lượng thông tin lớn từ nhiều nguồn khác nhau, cung cấp cái nhìn toàn diện và nhanh chóng về xu hướng giá cả, từ đó giảm thiểu rủi ro và tối ưu hóa lợi nhuận.

Mục tiêu của việc áp dụng Big Data vào dự đoán giá bất động sản là cung cấp các dự báo chính xác và kịp thời về biến động giá, giúp các bên liên quan như nhà đầu tư, ngân hàng, chính phủ và người mua nhà đưa ra các quyết định sáng suốt. Ngoài ra, nó còn nhằm cải thiện tính minh bạch của thị trường, hỗ trợ quy hoạch và phát triển đô thị bền vững, và nâng cao khả năng cạnh tranh của các doanh nghiệp trong ngành bất động sản.

1.3. Members and Role Assignments

* Lê Đông Anh Kiệt: Nhóm trưởng
  + Tiền xử lý dữ liệu
  + Hỗ trợ phân tích, trực quan hóa dữ liệu
  + Huấn luyện mô hình
* Trần Tuấn Kiệt:
  + Hỗ trợ phân tích, trực quan hóa dữ liệu
  + Hỗ trợ huấn luyện, đánh giá mô hình
  + Báo cáo thuyết trình
* Ngô Văn Lâu:
  + Báo cáo dự án
  + Hỗ trợ thu thập dữ liệu
  + Đánh giá mô hình
* Ngân Hoàng Huy:
  + Hỗ trợ báo cáo dự án
  + Hỗ trợ thu thập dữ liệu
  + Phân tích, đánh giá mô hình
* La Hồng Lộc:
  + Hỗ trợ báo cáo dự án
  + Hỗ trợ thu thập dữ liệu
  + Hỗ trợ huấn luyện mô hình
* Nguyễn Đặng Phú Mẫn:
  + Hỗ trợ báo cáo thuyết trình
  + Hỗ trợ phân tích, trực quan hóa dữ liệu

1.4. Schedule and Milestones

01/06 - 24/06 – Thu thập dữ liệu nhà đất từ Kaggle và Lianjia

18/06 - 07/07 – Tiền xử lý và làm sạch dữ liệu

25/06 - 07/07 – Xây dựng mô hình dự đoán

30/06 - 05/08 – Phân tích, trực quan hóa dữ liệu

08/07 – 14/08 – Huấn luyện mô hình dự đoán

01/08 – 14/08 – Viết báo cáo thống kê

2. Project Execution

2.1. Simulated Scenario Description

Trong những năm gần đây, thị trường bất động sản Bắc Kinh đã trải qua những biến động lớn về giá, có ảnh hưởng đáng kể đến nền kinh tế xã hội. Dự án này dựa trên một bộ dữ liệu giá nhà đất Bắc Kinh nhằm phát triển một mô hình phân tích để dự đoán giá bất động sản ở Bắc Kinh, sử dụng các thuật toán máy học để phân tích dữ liệu từ các giao dịch bất động sản và cung cấp cái nhìn sâu sắc hơn về xu hướng giá, giúp các nhà đầu tư và người tiêu dùng đưa ra các quyết định chính xác nhất.

2.2. Datasets Selection and Description

Lựa chọn Dữ liệu:

Dataset Selection: Dữ liệu được sử dụng trong dự án này bao gồm 318,851 bản ghi giao dịch bất động sản từ Bắc Kinh, Trung Quốc. Những bản ghi này bao gồm nhiều thuộc tính khác nhau cung cấp những thông tin quan trọng về động thái của thị trường bất động sản ở Bắc Kinh. Mỗi thuộc tính đều đóng vai trò quan trọng trong việc hiểu rõ giá bất động sản và xu hướng thị trường.

Attributes: Bộ dữ liệu bao gồm 26 thuộc tính, bao gồm các thông tin cơ bản như ID giao dịch, tọa độ vị trí, ngày giao dịch, giá bất động sản, diện tích, chi tiết phòng, loại tòa nhà, thời gian xây dựng, điều kiện sửa chữa, tiện nghi, thông tin quận huyện, và giá trung bình của cộng đồng.

Objective: Mục tiêu chính của việc lựa chọn bộ dữ liệu này là phát triển một mô hình dự đoán giá bất động sản ở Bắc Kinh. Bằng cách sử dụng các thuật toán máy học và kỹ thuật phân tích dữ liệu, mục tiêu là cung cấp hiểu biết chi tiết về xu hướng giá, hỗ trợ ra quyết định cho nhà đầu tư và người tiêu dùng, và nâng cao sự hiểu biết tổng thể về thị trường bất động sản ở Bắc Kinh.

Data Cleaning: Trước khi phân tích, bộ dữ liệu đã trải qua các bước tiền xử lý để giải quyết các giá trị thiếu, xử lý ngoại lệ, và đảm bảo chất lượng dữ liệu. Các quy trình làm sạch đã được thực hiện để chuẩn bị dữ liệu cho việc mô hình hóa, đảm bảo rằng nó chính xác và phù hợp cho phân tích.

Significance: Bộ dữ liệu được chọn là quan trọng để đánh giá các yếu tố ảnh hưởng đến giá bất động sản ở Bắc Kinh. Bằng cách xem xét các thuộc tính như vị trí, đặc điểm bất động sản, tiện nghi, và xu hướng thị trường, các bên liên quan có thể thu được những thông tin giá trị giúp hỗ trợ ra quyết định chiến lược và phân tích thị trường.

Mô tả thuộc tính Dữ liệu:

|  |  |  |
| --- | --- | --- |
| Tên thuộc tính | Mô tả | Kiểu dữ liệu |
| url | Đường dẫn để lấy data. | string |
| id | Mã định danh cho mỗi giao dịch bất động sản. | string |
| Lng | Kinh độ của vị trí nhà đất theo bản đồ thế giới. | float |
| Lat | Vĩ độ của vị trí nhà đất theo bản đồ thế giới. | float |
| Cid | Mã định danh cộng đồng. | string |
| tradeTime | Ngày diễn ra giao dịch bất động sản. | datetime |
| DOM | Thời gian mà bất động sản tồn tại trên thị trường (Days on market). | integer |
| followers | Số lượng người theo dõi giao dịch bất động sản. | integer |
| totalPrice | Tổng giá thành của bất động sản (CNY). | float |
| price | Đơn giá của bất động sản tính trên mỗi đơn vị diện tích (CNY/m²). | float |
| square | Diện tích của bất động sản (m²). | float |
| livingRoom | Số lượng phòng khách trong nhà. | integer |
| drawingRoom | Số lượng phòng vẽ hoặc phòng khách thứ hai trong bất động sản. | integer |
| kitchen | Số lượng phòng bếp trong nhà. | integer |
| bathRoom | Số lượng phòng tắm trong nhà. | integer |
| floor | Độ cao và số tầng của tòa nhà bất động sản. | string |
| buildingType | Loại hình tòa nhà (ví dụ: thiết kể phằng, cao ốc). | integer |
| constructionTime | Năm xây dựng của tòa nhà bất động sản. | integer |
| renovationCondition | Điều kiện tu sửa của bất động sản (ví dụ: tu sửa thô, đơn giản, tu sửa cứng). | integer |
| buildingStructure | Cấu trúc của tòa nhà (ví dụ: khung thép, gỗ, gạch, bê tông). | integer |
| ladderRatio | Tỷ lệ giữa số người dân và số thang bộ/thang máy trong tòa nhà. | float |
| elevator | Có hay không thang máy trong nhà đất. | boolean |
| fiveYearsProperty | Tình trạng sở hữu đất trong 5 năm. | boolean |
| subway | Gần hay không gần tàu điện ngầm. | boolean |
| district | Quận, huyện, khu vực tại Bắc Kinh. | integer |
| communityAverage | Giá trung bình của cộng đồng (CNY). | float |

2.3. Data Ingestion Pipeline

Nhập các thư viện và thiết lập môi trường: Bước đầu tiên trong quá trình thu thập dữ liệu liên quan đến việc thiết lập môi trường và nhập các thư viện cần thiết. Điều này bao gồm các thư viện cho việc xử lý dữ liệu, trực quan hóa và máy học.

Đọc dữ liệu: Bước tiếp theo liên quan đến việc đọc bộ dữ liệu bằng thư viện Pandas. Vì tập tin chứa các ký tự tiếng Trung, nên mã hóa 'GBK' được sử dụng để đảm bảo việc đọc dữ liệu đúng cách.

Quá trình thiết lập và thu thập này là rất quan trọng để chuẩn bị dữ liệu cho phân tích và mô hình hóa tiếp theo. Nó đảm bảo rằng các công cụ cần thiết có sẵn và dữ liệu được đọc đúng vào định dạng phù hợp cho việc xử lý và phân tích.

2.4. Data Transformation Processing

2.4.1 Xử lý dữ liệu

* Xử lý thông tin:
* Trong dataset có một số hàng không phải sai thông tin mà lỗi nhập liệu nhầm thông tin thuộc tính cho nhau. Ví dụ: cột buildingType hiển thị thông tin ladderRatio, bathRoom hiển thị thông tin constructionTime, ... Xác định và xử lý mà không cần bỏ những thông tin ấy.
* Chuyển các giá trị được gọi là rỗng như “”, “#NAME?”, “nan”, “ “, “未知” thành giá trị rỗng kiểu dữ liệu numpy.nan để dễ xử lý.

Hình 1: Một số hàng nhập liệu thông tin nhầm thuộc tính so với các thông tin khác.A screenshot of a computer

Description automatically generated

Hình 2: Các hàng nhập liệu đúng thông tin so với Hình 1.A screenshot of a computer

Description automatically generated

* Xử lý cột:
* Tạo cột floorType để chứa phần không phải số từ cột floor.
* Cập nhật cột floor để chỉ chứa chỉ số tầng từ cột floor.
* Ánh xạ các giá trị văn bản trong cột floorType thành các chỉ số phân loại kiểu nhà tầng.
* Thay thế các giá trị trong cột buildingStructure với các giá trị số cụ thể.
* Tách các giá trị từ cột tradeTime thành ba cột mới: tradeYear, tradeMonth, và tradeDay, tất cả đều được chuyển đổi thành kiểu số nguyên.
* Thuộc tính DOM có ảnh hưởng cao đến giá thành nhà đất nên nhóm cần xử lý đặc biệt:
* Đầu tiên là xác định các đặc trưng có liên kết cao với nó, đó là: tradeYear, district và followers.
* Vì followers là một biến số, nên không tiện dụng để gộp, do đó chúng tôi chuyển nó thành một biến kiểu chuỗi - follow\_type.
* Đánh giá các giá trị rỗng trong DOM bằng cách tính trung bình của dữ liệu có cùng giá trị với chúng trong ba tính năng tradeYear, district, follow\_type tương ứng.
* Với việc xử lý giá trị khuyết tại các thuộc tính còn lại, thuộc tính có kiểu dữ liệu định danh sẽ thay thế bằng giá trị phổ biến nhất (mode). Còn các kiểu còn lại, tùy thuộc vào độ lệch chuẩn của phân phối dữ liệu sẽ thay thế thành giá trị trung bình (mean) hoặc trung vị (median)

2.4.2. Biến đổi dữ liệu

* Xác định lại kiểu dữ liệu phù hợp cho từng thuộc tính: thông tin năm tháng ngày nên theo số, thông tin phân loại không là số thực
* Chuẩn hóa dữ liệu sử dụng kỹ thuật Min-Max Scaler: là một phương pháp tiền xử lý dữ liệu thường được sử dụng trong máy học (machine learning). Nó chuẩn hóa dữ liệu bằng cách đưa tất cả các giá trị về một khoảng xác định, thường là [0, 1]. Công thức chuẩn hóa của MinMaxScaler là:

A black and white math equation

Description automatically generated

* Trong đó:
* X là giá trị ban đầu.
* Xmin và Xmax lần lượt là giá trị nhỏ nhất và lớn nhất của dữ liệu.
* Xử lý giá trị ngoại lệ bằng Interquartile Range

**Interquartile Range (IQR)** là một thước đo thống kê được sử dụng để đánh giá mức độ phân tán của một tập dữ liệu, đặc biệt là mức độ phân tán của các giá trị trung bình. IQR được tính bằng cách lấy giá trị chênh lệch giữa phân vị thứ 3 (Q3) và phân vị thứ 1 (Q1) của dữ liệu.

Cụ thể:

* **Q1 (Phân vị thứ 1)**: là giá trị tại vị trí 25% trong tập dữ liệu đã được sắp xếp theo thứ tự tăng dần. Nói cách khác, 25% dữ liệu nằm dưới Q1.
* **Q3 (Phân vị thứ 3)**: là giá trị tại vị trí 75% trong tập dữ liệu đã được sắp xếp theo thứ tự tăng dần. Tức là 75% dữ liệu nằm dưới Q3.

**Công thức tính IQR:**

**IQR =Q3−Q1**

**IQR** được sử dụng để xác định các giá trị ngoại lệ (outliers) trong tập dữ liệu. Một cách phổ biến để phát hiện outliers là kiểm tra các giá trị nằm ngoài khoảng **[Q1 - 1.5 \* IQR, Q3 + 1.5 \* IQR]**.

2.5. Data Query and Insight

* Dữ liệu có hình dạng chuẩn mà không đều, với đỉnh nằm ở khoảng giữa, có vẻ như phân bố lệch trái. Điều này cho thấy đa số bất động sản có giá thấp đến trung bình, với ít bất động sản có giá cao.

A graph of a distribution of prices

Description automatically generatedA graph of a distribution of prices

Description automatically generated

Hình 3: Biểu đồ phân phối giá thành nhà đất.

A graph of a graph with different colored squares

Description automatically generated with medium confidenceA chart with numbers and letters

Description automatically generated with medium confidence

Hình 4: Biểu đồ nhiệt tương quan. Hình 5: Biểu đồ hộp tương quan theo giá thành.

A graph of a line

Description automatically generated with medium confidence

Hình 6: Biểu đồ biến động giá cả theo năm

* **Hình 3:** Thể hiện một mối quan hệ tăng dần, nhưng không hoàn toàn tuyến tính. Khi diện tích tăng, giá cũng tăng theo nhưng với tốc độ không đồng đều.
* **Hình 4:** Thể hiện dữ liệu phân tán rộng, cho thấy có sự đa dạng trong các loại bất động sản với các mức giá khác nhau. Thuộc tính **square** đều có tương quan cao với **totalPrice**, điều này là dễ hiểu vì tổng giá cả thường tăng lên diện tích của bất động sản.
* **Hình 5:** Có cột livingRoom, drawingRoom và communityAverage cũng cho thấy mối tương quan mạnh, cho thấy rằng những bất động sản có nhiều phòng khách hoặc nằm trong cộng đồng có giá trị trung bình cao thường có giá bán tổng thể cao. Xu hướng giá tăng dần: giá nhà đất đã tăng qua các năm từ 2010 đến 2018.
* **Hình 6**: Biến động giá chênh lệch: Các điểm dữ liệu nằm ngoài các "whiskers" của boxplot (các outliers) cho thấy có một số giao dịch với giá bất động sản cao bất thường so với mức trung bình của tháng đó, điều này có thể do các giao dịch bất động sản cao cấp hoặc các yếu tố khác ảnh hưởng.

2.6. Prediction Methods:

Hồi quy tuyến tính (Linear Regression):

* Là một mô hình máy học đơn giản nhất dùng để dự đoán một giá trị số (regression) dựa trên mối quan hệ tuyến tính giữa biến đầu vào (features) và biến đầu ra (target)
* Mô hình này cố gắng tìm một đường thẳng (hàm tuyến tính) sao cho tổng bình phương sai số giữa dự đoán của mô hình và giá trị thực tế là nhỏ nhất. Công thức của nó là:

Cây quyết định (Decision tree):

* Là một mô hình máy học dự đoán giá trị số bằng cách phân chia dữ liệu thành các nhánh hoặc nút dựa trên các điều kiện trên các đặc trưng đầu vào.
* Mô hình chia dữ liệu thành các nhánh dựa trên các điều kiện có thể làm giảm thiểu lỗi dự đoán (thường được đo bằng phương sai). Quá trình phân nhánh này tiếp tục cho đến khi đạt được một điều kiện dừng nào đó, như khi đạt đến độ sâu tối đa của cây hoặc khi mỗi lá chỉ chứa một số mẫu nhất định.

Gradient Boosting:

* Là một thuật toán tối ưu hóa tăng cường (Ensemble Learning) phổ biển, dùng để xây dựng các mô hình dự đoán yếu kết hợp thành một mô hình máy học mạnh mẽ.
* Thuật toán bắt đầu huấn luyện dự đoán yếu, thường là Decision Tree, sau đó tiếp tục thêm các mô hình mới, mỗi lần được huấn luyện sẽ khắc phục các dự đoán sai trước đó. Quá trình được lặp lại nhiều lần, tăng cường hiệu suất cho đến khi đạt được mức độ hiểu quả mong muốn hoặc số lượng mô hình được thêm đạt đến giới hạn.
* Khác với Gradient Descent, một thuật toán tối ưu hóa để tìm ra giá trị tối thiểu của hàm mất mát (loss function). Bằng cách cập nhật tham số liên tục cho đến khi giá trị theo hướng ngược lại làm giảm thiểu hàm mất mát, các mô hình trong bài báo cáo đều sử dụng thuật toán này.

Mạng nơ-ron nhân tạo (Artificial Neural Network – ANN):

* Là một mô hình được thiết kế dựa trên cách hoạt động của bộ não con người.
* ANN bao gồm nhiều lớp nơ-ron, trong đó mỗi nơ-ron nhận đầu vào, thực hiện các phép tính (thường là kết hợp tuyến tính và hàm kích hoạt) và truyền kết quả tới các nơ-ron ở lớp tiếp theo. Quá trình huấn luyện điều chỉnh trọng số của các liên kết giữa các nơ-ron sao cho mạng học được cách dự đoán kết quả một cách chính xác.
* Loại ANN được sử dụng trong dự án là mô hình Perceptron đa lớp (MultiLayer Perceptron – MLP), với kiến trúc tiến thẳng (Feedforward) và kết nối đầy đủ, dữ liệu chạy từ lớp đầu vào đến lớp đầu ra mà không lặp lại và mỗi nơ-ron, trong một lớp được kết nối với mọi nơ-ron trong lớp tiếp theo, cho nên có mạng lưới nơ-ron kết nối một cách dày đặc.

Đánh giá Hiệu suất:

* Áp dụng các chỉ số đánh giá vững chắc để đánh giá độ chính xác của mô hình, như RMSE (Lỗi bình phương trung bình gốc) và MAE (Lỗi tuyệt đối trung bình). Xác thực các mô hình bằng cách sử dụng kiểm tra chéo để đảm bảo khả năng tổng quát và ngăn ngừa việc quá khớp.

3. Results

3.1. Data Ingestion Scripts and Code

df = pd.read\_csv(‘rawdata.csv', encoding='gbk', low\_memory=False)

* Nhóm sử dụng một hàm trong thư viện Pandas là pd.read\_csv() để đọc dữ liệu từ một tập tin rawdata.csv và chuyển tập đó thành DataFrame (một cấu trúc dữ liệu dạng bảng với các hàng và cột).
* Tham số encoding='gbk' chỉ định cách giải mã ký tự trong tập tin được đọc, GBK được sử dụng để mã hóa các ký tự tiếng Trung Quốc để tránh lỗi đọc tập tin.

Tham số low\_memory=False được thiết kế để tối ưu hóa việc đọc tập tin lớn, thư viện sẽ không tự đoán kiểu dữ liệu của cột và sử dụng nhiều bộ nhớ hơn để xác định kiểu dữ liệu chính xác hơn. Như vậy giúp giảm thiểu khả năng xảy ra lỗi kiểu dữ liệu không mong muốn khi xử lý tập CSV lớn hoặc phức tạp.

3.2. Data Transformation Scripts and Code

bins = [-1, 0, 5, 10, 20, 50, 9999]

df['follow\_type'], b = pd.cut(x=df['followers'], bins=bins, right=True, retbins=True)

df['DOM'].fillna(df.groupby(['district', 'tradeYear', 'follow\_type'])['DOM']

.transform('mean'), inplace=True)

df['DOM'].fillna(df['DOM'].mean(), inplace=True)

df.drop('follow\_type', axis=1, inplace=True)

* bins là danh sách các ngưỡng giá trị phân chia dữ liệu cột followers, mỗi khoảng giữa các giá trị là một nhóm.
* Hàm pd.cut() phân chia giá trị trong cột vào khoảng trên, trả về các bins mới và tạo cột mới follow\_type lưu kết quá phân loại.
* Bắt đầu điền giá trị thiếu vào cột DOM dựa trên các đặc trưng có liên kết cao và tính giá trị trung bình cho mỗi nhóm, tiếp theo cập nhật trực tiếp vào df.
* Nếu sau khi điền theo nhóm vẫn còn giá trị rỗng thì thế bằng giá trị trung bình cột DOM, cuối cùng là xóa cột follow\_type khỏi df, đảm bảo dữ liệu được chuẩn bị và làm sạch một cách chính xác nhất.

Q1 = df[['totalPrice', 'square']].quantile(0.25)

Q3 = df[['totalPrice', 'square']].quantile(0.75)

IQR = Q3 - Q1

outliers = ((df[['totalPrice', 'square']] < (Q1 - 1.5 \* IQR)) |

(df[['totalPrice', 'square']] > (Q3 + 1.5 \* IQR))).any(axis=1)

df = df[~outliers]

* Đây là kỹ thuật phổ biến phát hiện và loại bỏ các giá trị ngoại lai (outliers)
  + Q3 là giá trị phân vị 0.75, nơi 75% dữ liệu có giá trị thấp hơn hoặc bằng nó.
  + IQR (Interquantile Range) là khoảng cách giữa Q1 và Q3, đo độ phân tán của dữ liệu.
  + Q1 là giá trị phân vị 0.25, nơi 25% dữ liệu có giá trị thấp hơn hoặc bằng nó.
* Ta xác định outliers dựa trên Rule of Thumb, nơi mà một giá trị được xem là ngoại lai nó nằm ngoài khoảng [Q1 – 1.5 \* IQR, Q3 + 1.5 \* IQR]. Kết quả đưa ra là chuỗi các Boolean cho mỗi hang chứa ít nhất một outlier.
* Cập nhật lại df sử dụng toán tử ~ lấy những hàng không là outlier.

3.3. Description and Sample of Transformed Datasets

A screenshot of a table

Description automatically generatedHình 7: Mẫu tập dữ liệu đã biến đổi.

3.4. Data Visualization of Query Results

sns.histplot(df['totalPrice'], kde=True, log\_scale=True)

plt.title('Distribution of Prices')

plt.show()

A graph of a distribution of prices

Description automatically generated

sns.heatmap(df.corr(), annot=False, cmap='coolwarm', linewidths=1)

plt.title('Correlation Matrix')

plt.show()

A graph of a graph with different colored squares

Description automatically generated with medium confidence

df = df.drop(['district\_name'], axis=1, errors='ignore')

features = df.corr()[['totalPrice']].sort\_values(by=['totalPrice'], ascending=False)

plt.figure(figsize=(5, 10))

sns.heatmap(features, cmap='rainbow', annot=True, annot\_kws={"size": 16}, vmin=-1)

plt.show()

A chart of a number of objects

Description automatically generated with medium confidence

dt = df.groupby(['tradeYear', 'tradeMonth', 'tradeDay'])['totalPrice'].mean().reset\_index()

dt = dt[dt['tradeYear'] >= 2010]

date\_df = pd.DataFrame({

'year': dt['tradeYear'],

'month': dt['tradeMonth'],

'day': 1

})

dt['tradeTime'] = pd.to\_datetime(date\_df)

fig, ax = plt.subplots(figsize=(20, 10))

fig.suptitle('House Prices Over Time', fontsize=20)

palette = sns.color\_palette("rainbow", n\_colors=dt['tradeYear'].nunique())

sns.boxplot(data=dt, x='tradeTime', y='totalPrice', saturation=0.7, ax=ax, palette=palette)

plt.xticks(np.arange(0, len(dt['tradeTime'].unique()), 12), labels=range(2010, 2018+1))

ax.grid(True)

A graph of a line

Description automatically generated with medium confidence

3.5. Experimental results

A comparison of a graph

Description automatically generated with medium confidence

* Random Forest là mô hình tốt nhất trong tất cả mô hình với kết quả MAE, MSE, RMSE thấp nhất và R2 cao nhất.
* MLP, Decision Tree và Gradient Boosting có kết quả tương đối ổn định với hiệu suất tốt về mặt sai số thấp.
* Linear Regression là mô hình cho thấy tỷ lệ lỗi cao nhất và hiệu suất thấp nhất.

A blue and red dotted graph

Description automatically generated

* Linear Regression có sự khác biệt đáng kể so với đường 𝑦 = 𝑥, cho thấy dấu hiệu của sự thiếu khớp và hiệu suất dự đoán kém.
* Decision Tree và Gradient Boosting gần hơn với đường thẳng nhưng vẫn thể hiện sự phân tán, cho thấy mức độ khớp vừa phải, xử lý ảnh hưởng của đặc trưng tốt hơn.
* Random Forest và MLP thể hiện sự khớp tốt nhất với đường 𝑦 = 𝑥, cho thấy dự đoán xuất sắc trong phạm vi.

A graph of a blue line

Description automatically generated with medium confidence

* Tất cả biểu đồ càng hẹp và tập trung xung quanh số 0 càng tốt, vì điều này cho thấy sai số nhỏ và không thiên vị.
* Linear Regression, Gradient Boosting và Decision Tree hiển thị phân bố rộng và ít tập trung, cho thấy sự phân tán trong lỗi và thiên vị tiềm ẩn.
* Random Forest và MLP thể hiện phân bố đỉnh cao và chính xác xung quanh số 0, cho thấy độ chính xác và nhất quán tốt của mô hình.

Kết quả đánh giá mô hình:

A screenshot of a graph

Description automatically generated

4. Projected Impact

4.1. Accomplishments and Benefits

**Thành tựu:**

Xử lý Dữ liệu Toàn diện: Đã thành công trong việc làm sạch và tích hợp một tập dữ liệu đa dạng, xử lý các giá trị thiếu và ngoại lệ để đảm bảo chất lượng dữ liệu cao. Biến đổi dữ liệu thô thành định dạng cấu trúc phù hợp cho phân tích và mô hình hóa.

Kỹ thuật và Lựa chọn Đặc trưng: Tạo các đặc trưng mới từ dữ liệu hiện có, chẳng hạn như chia ngày giao dịch thành các thành phần riêng biệt. Lựa chọn và mã hóa các đặc trưng liên quan, tối ưu hóa chúng cho các thuật toán máy học.

Phát triển Mô hình: Xây dựng và kiểm tra nhiều mô hình máy học khác nhau.

Và các phương pháp tập hợp. Tinh chỉnh các siêu tham số để cải thiện hiệu suất mô hình và đảm bảo dự đoán chính xác.

Trực quan hóa và Nhận thức: Tạo các trực quan hóa để tiết lộ những hiểu biết về xu hướng giá cả, tầm quan trọng của các đặc trưng và phân bố địa lý. Tạo các biểu đồ dễ sử dụng để truyền đạt các phát hiện đến các bên liên quan.

**Lợi ích:**

Dự Đoán Giá Cả Chính Xác: Cung cấp các dự đoán giá nhà đất đáng tin cậy, hỗ trợ trong việc ra quyết định đầu tư và phân tích thị trường.

Hiểu Biết Thị Trường Cải Thiện: Cung cấp những hiểu biết về xu hướng thị trường nhà ở và các yếu tố chính ảnh hưởng đến giá cả, giúp các bên liên quan đưa ra quyết định thông minh.

Quyết Định Dựa Trên Dữ Liệu: Cho phép các nhà phát triển bất động sản, nhà đầu tư và các nhà hoạch định chính sách sử dụng các chiến lược dựa trên dữ liệu để đạt được kết quả tốt hơn.

Khả Năng Mở Rộng và Tái Tạo: Phát triển các quy trình xử lý và mô hình hóa dữ liệu có thể mở rộng và tái tạo, tạo điều kiện cho các cải tiến và thích ứng trong tương lai.

4.2. Future Improvements

* **Xử lý dữ liệu thiếu và biến đổi dữ liệu:**
* Tích hợp các kỹ thuật tiên tiến hơn để xử lý các giá trị thiếu, đặc biệt là trong cột DOM bằng cách sử dụng các mô hình dự đoán hoặc các kỹ thuật thống kê hiện đại hơn.
* Thực hiện việc chuẩn hóa và chuyển đổi các biến phân loại với các phương pháp mã hóa tiên tiến như Label Encoding hoặc Target Encoding, thay vì chỉ sử dụng Min-Max Scaling hiện tại.
* **Cải thiện mô hình máy học:**
* Nghiên cứu và thử nghiệm các thuật toán máy học khác, chẳng hạn như các biến thể của XGBoost để cải thiện độ chính xác của mô hình.
* Kết hợp nhiều mô hình (ensemble) để tạo ra một mô hình tổng hợp với hiệu suất tốt hơn.
* **Phân tích thêm các yếu tố ảnh hưởng đến giá nhà** **đất:**
* Tăng cường phân tích các yếu tố khác như vị trí, tiện ích, cơ sở hạ tầng xung quanh, và điều kiện thị trường để xây dựng mô hình dự đoán toàn diện hơn.
* Sử dụng các phương pháp phân tích không gian địa lý để đánh giá sự ảnh hưởng của vị trí địa lý đối với giá nhà đất.
* **Áp dụng các phương pháp học sâu:**
* Khám phá việc sử dụng các mô hình học sâu (Deep Learning) như Mạng Nơ-ron Tích chập (Convolutional Neural Networks - CNN) hoặc Mạng Nơ-ron Hồi quy (Recurrent Neural Networks - RNN) để dự đoán giá nhà đất từ dữ liệu ảnh và chuỗi thời gian nếu có.

5. Team Member Review and Comment

|  |
| --- |
| <ATTACH A TEAM PICTURE HERE> |

|  |  |
| --- | --- |
| NAME | REVIEW and COMMENT |
| Lê Đông Anh Kiệt |  |
| Trần Tuấn Kiệt |  |
| Ngô Văn Lâu |  |
| La Hồng Lộc |  |
| Ngân Hoàng Huy |  |
| Nguyễn Đặng Phú Mẫn |  |

6. Instructor Review and Comment

|  |  |  |
| --- | --- | --- |
| CATEGORY | SCORE | REVIEW and COMMENT |
| IDEA | \_\_/10 |  |
| APPLICATION | \_\_/30 |  |
| RESULT | \_\_/30 |  |
| PROJECT MANAGEMENT | \_\_/10 |  |
| PRESENTATION & REPORT | \_\_/20 |  |
| TOTAL | \_\_/100 |  |