TRƯỜNG PHÂN HIỆU ĐẠI HỌC THỦY LỢI

ĐỀ TÀI: DỰ ĐOÁN GIÁ NHÀ BẤT ĐỘNG SẢN

SINH VIÊN THỰC HIỆN: NGUYỄN NGỌC BẢO

MÃ SỐ SINH VIÊN: 2051067522

GIẢNG VIÊN HƯỚNG DẪN: Ths. Vũ Thị Hạnh

TP HCM, NGÀY 23 THÁNG 10 NĂM 2025

LỜI CẢM ƠN

Lời cảm ơn đầu tiên của chúng em được gửi đến các thầy cô giảng dạy trong
trường Phân hiệu ĐH Thủy Lợi. Chúng em muốn gửi lời cảm ơn sâu sắc nhất đến những, người đã truyền dạy cho chúng em những kiến thức quý báu, giúp chúng em phát triển, và tiến bộ trong cuộc sống.
Những giáo viên tận tâm và nhiệt tình đã luôn sẵn sàng, giúp đỡ chúng em trong suốt quá trình học tập.Đặc biệt, chúng em muốn gửi lời cảm ơn đến ThS. Vũ Thị Hạnh đã truyền đạt những kiến thức bổ ích và kinh nghiệm thực tiễn giúp chúng em hiểu rõ hơn về đề tài và hoàn thành bài luận án một cách tốt nhất.
Cuối cùng, chúng em xin chân thành cảm ơn các thầy cô trong trường ĐH Thủy Lợi đã tạo điều kiện tốt nhất cho chúng em trong suốt thời gian học tập tại trường. Chúng em hy vọng sẽ có cơ hội được tiếp tục học tập và phát triển bản thân
trong tương lai.

TÓM TẮT

Báo cáo này tập trung vào nghiên cứu các mô hình để dự đoán giá nhà trên dữ liệu bảng. Thị trường bất động sản là một lĩnh vực phức tạp và năng động; việc dự đoán giá nhà chính xác có ý nghĩa quan trọng đối với người mua, người bán, nhà đầu tư và các cơ quan định giá, giúp tối ưu hóa các quyết định tài chính và chiến lược kinh doanh.

Chúng tôi sử dụng một tập dữ liệu chi tiết bao gồm thông tin về các đặc trưng của căn nhà như diện tích, số lượng phòng ngủ và phòng tắm, tuổi của căn nhà, vật liệu xây dựng, cũng như các yếu tố liên quan đến vị trí như khu vực, khoảng cách đến các tiện ích công cộng, và các chỉ số kinh tế xã hội của khu phố. Mục tiêu của chúng tôi là xây dựng một mô hình hồi quy hiệu quả để dự đoán giá trị thị trường ước tính của một căn nhà dựa trên các yếu tố đầu vào này.

Báo cáo này bao gồm một phân tích chi tiết về việc tiền xử lý dữ liệu (xử lý dữ liệu khuyết, biến định tính), lựa chọn và đánh giá mô hình hồi quy cùng với việc giải thích sâu hơn về cách các đặc trưng ảnh hưởng đến dự đoán giá. Kết quả được thảo luận cung cấp một cái nhìn tổng quan về hiệu suất của mô hình và các yếu tố quan trọng nhất trong việc dự đoán giá trị bất động sản.

MỤC LỤC

CHƯƠNG 1 – GIỚI THIỆU VÀ TỔNG QUAN VỀ ĐỀ TÀI

1.1. Giới thiệu về bất động sản tại thành phố lowa:

1.1.1. Định nghĩa:

-Bất động sản là tài sản bao gồm đất đai và những cấu trúc được xây dựng hoặc gắn liền vĩnh viễn với đất đai đó, như nhà ở, tòa nhà thương mại và cơ sở hạ tầng. Trong bối cảnh của báo cáo này, trọng tâm là nhà ở những tài sản được sử dụng làm nơi cư trú.

-Việc định giá và giao dịch bất động sản là một quy trình phức tạp, chịu ảnh hưởng của vô số yếu tố, từ các đặc điểm vật lý của tài sản đến điều kiện kinh tế vĩ mô.

1.1.2. Tổng Quan về Thị Trường Nhà Ở tại Thành phố Iowa:

-Thành phố Iowa là một trung tâm văn hóa và giáo dục quan trọng, là nơi đặt trụ sở của Đại học Iowa Sự hiện diện của trường đại học này là một động lực chính định hình thị trường bất động sản tại đây.

-Đặc điểm thị trường: Thị trường nhà ở Iowa City nổi tiếng với sự ổn định và nhu cầu thuê và mua cao do số lượng lớn sinh viên, giảng viên và nhân viên y tế

-Ảnh hưởng của Vị trí: Giá nhà có xu hướng cao hơn ở các khu vực gần khuôn viên trường đại học và trung tâm thành phố

-Mục tiêu dự đoán: Việc dự đoán giá nhà chính xác tại Iowa City là rất quan trọng để giúp người mua và nhà đầu tư đưa ra quyết định sáng suốt, đặc biệt khi cân nhắc các yếu tố như: khoảng cách đến trường học, tuổi và tình trạng của tài sản, và các chỉ số kinh tế địa phương.

-Thị trường nhà ở tại Iowa City là một ví dụ điển hình cho thấy sự tương tác phức tạp giữa các đặc trưng vật lý của căn nhà (diện tích, số phòng, chất lượng xây dựng) và các yếu tố bên ngoài (vị trí, điều kiện kinh tế, cầu từ sinh viên) trong việc xác định giá trị cuối cùng.

1.1.3. Tầm quan trọng của việc dự đoán:

- Định giá công bằng:

+Người Bán: Dự đoán giúp họ xác định mức giá niêm yết tối ưu, đảm bảo bán được tài sản trong thời gian ngắn nhất mà vẫn đạt được lợi nhuận cao nhất.

+Người Mua: Dự đoán cung cấp một điểm tham chiếu khách quan để đánh giá liệu giá chào bán có hợp lý hay không, giúp tránh việc mua với giá quá cao

-Ra quyết định Tài chính Cá nhân: Hỗ trợ người mua ước tính khoản vay thế chấp và lập kế hoạch tài chính hiệu quả hơn.

-Phân tích Lợi nhuận Đầu tư (Mô hình dự đoán là công cụ thiết yếu để nhà đầu tư ước tính giá trị tương lai của tài sản, từ đó xác định Lợi nhuận trên vốn đầu tư của các dự án mua, sửa chữa hoặc xây dựng mới.

-Quản lý Rủi ro: Giúp các nhà đầu tư nhận diện và định lượng rủi ro liên quan đến sự biến động giá trị tài sản trong các chu kỳ thị trường khác nhau.

- Lập kế hoạch Phát triển: Hỗ trợ các nhà phát triển xác định khu vực có nhu cầu cao và tiềm năng tăng giá trong tương lai để tập trung nguồn lực xây dựng.

1.2. Mục tiêu báo cáo:

1.2.1. Lý do chọn đề tài:

-Tính Ổn Định và Tác động của Đại học: Thị trường bất động sản Iowa City chịu ảnh hưởng mạnh mẽ và tương đối ổn định từ Đại học Iowa. Điều này tạo ra một môi trường lý tưởng để nghiên cứu các yếu tố ảnh hưởng đến giá nhà một cách có hệ thống, đồng thời cung cấp công cụ hữu ích cho cộng đồng địa phương (sinh viên, nhân viên, nhà đầu tư).

- Thiếu Công cụ Định giá Tự động: Mặc dù có các dịch vụ định giá truyền thống, nhưng vẫn cần một công cụ định giá tự động chính xác và kịp thời, đặc biệt quan trọng trong các giao dịch nhanh hoặc khi cần đánh giá danh mục lớn.

- Hỗ trợ Ra Quyết Định: Mô hình dự đoán giúp:

+Người Mua/Bán: Có cơ sở dữ liệu để đàm phán, tránh định giá sai lệch.

+Nhà Đầu tư: Xác định các tài sản bị định giá thấp hoặc định giá quá cao để tối ưu hóa lợi nhuận.

1.2.2. Mục tiêu nghiên cứu và phạm vi:

-Phân tích và Tiền xử lý Dữ liệu: Thực hiện phân tích khám phá dữ liệuvà các bước tiền xử lý cần thiết để chuẩn bị dữ liệu cho việc huấn luyện mô hình.

- Xây dựng và Huấn luyện Mô hình: Áp dụng và huấn luyện ít nhất ba mô hình hồi quy khác nhau (ví dụ: Linear Regression, Random Forest, Gradient Boosting) trên tập dữ liệu đã được tiền xử lý.

- Đánh giá Hiệu suất Mô hình: Đánh giá hiệu suất của các mô hình đã xây dựng bằng các chỉ số hồi quy tiêu để xác định mô hình có hiệu suất tốt nhất.

- Giải thích Đặc trưng Phân tích và giải thích tầm quan trọng của các đặc trưng đối với việc dự đoán giá, từ đó cung cấp hiểu biết sâu sắc về các yếu tố chi phối thị trường bất động sản Iowa City.

1.2.3. Ý nghĩa của đề tài:

-Nghiên cứu tạo ra một công cụ định giá và phân tích có ý nghĩa trực tiếp đối với các bên liên quan trong thị trường bất động sản:

+Tăng tính Minh bạch và Công bằng Giá cả: Xây dựng một Mô hình Định giá Tự động (AVM) khách quan giúp chuẩn hóa quá trình định giá. Điều này hỗ trợ người mua và người bán trong việc thương lượng và tránh tình trạng định giá quá cao (overpriced) hoặc quá thấp (underpriced).

+Hỗ trợ Quyết định Tài chính và Đầu tư:

Giúp Ngân hàng thẩm định giá trị tài sản thế chấp nhanh chóng và chính xác, giảm thiểu rủi ro cho vay, cung cấp cơ sở dữ liệu cho Nhà đầu tư để xác định các bất động sản tiềm năng và tối ưu hóa chiến lược mua, bán, hoặc cải tạo.

+Phân tích Thị trường Sâu sắc: Kết quả mô hình (đặc biệt là phân tích Tầm quan trọng của Đặc trưng) cung cấp cái nhìn sâu sắc về các yếu tố chi phối giá nhà tại Iowa City (ví dụ: ảnh hưởng của khoảng cách đến Đại học Iowa, diện tích, hay chất lượng vật liệu), giúp các bên hiểu rõ hơn về cơ chế thị trường địa phương.

CHƯƠNG 2 – MÔ TẢ VÀ KHÁM PHÁ BỘ DỮ LIỆU

2.1. Mô tả bộ dữ liệu:

2.1.1. Ngữ cảnh:

-Nghiên cứu này được đặt trong ngữ cảnh của Khoa học Dữ liệu (Data Science) nhằm giải quyết một bài toán kinh tế thực tiễn bằng cách áp dụng các mô hình Học máy (Machine Learning).

- Bối cảnh Địa lý: Nghiên cứu tập trung vào thị trường nhà ở tại Thành phố Iowa Hoa Kỳ, một thị trường có đặc thù ổn định và chịu ảnh hưởng lớn từ sự hiện diện của Đại học Iowa.
- Vấn đề: Xác định giá trị thị trường của nhà ở là một quá trình phức tạp và thường mang tính chủ quan. Nhu cầu đặt ra là cần một công cụ định giá tự động khách quan và chính xác.
- Phương pháp: Sử dụng dữ liệu bảng về các giao dịch bán nhà, nghiên cứu này áp dụng các mô hình hồi quy tiên tiến để dự đoán biến mục tiêu là Giá bán
- Mục tiêu: Xây dựng một mô hình dự đoán hiệu suất cao, đồng thời giải thích được các yếu tố nào là quan trọng nhất trong việc định giá nhà tại Iowa City.
- Ngữ cảnh này định hình toàn bộ phương pháp luận và mục tiêu của báo cáo, biến nó thành một dự án áp dụng công nghệ tiên tiến (để tạo ra giá trị kinh tế

2.1.2. Các đặc trưng trong bộ dữ liệu

Bộ dữ liệu gồm file train.csv và test.csv, describe, sample, với các đặc trưng được mô tả

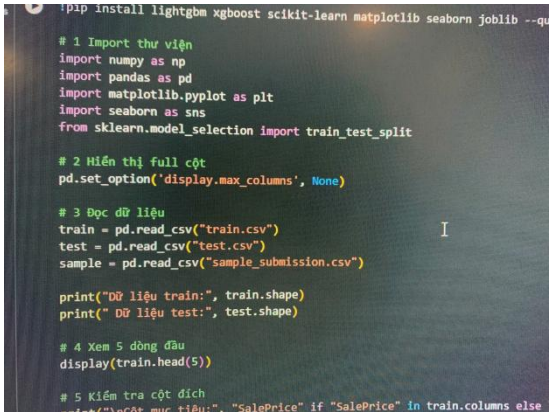
2.2. Khám phá bộ dữ liệu:

-SalePrice - giá bán bất động sản tính bằng đô la. Đây là biến mục tiêu mà bạn đang cố gắng dự đoán.
-MSSubClass : Lớp xây dựng
-MSZoning : Phân loại phân vùng chung
-LotFrontage : Chiều dài tuyến tính của đường phố được kết nối với bất động sản
-LotArea : Diện tích lô đất tính bằng feet vuông
-Đường phố : Loại đường đi vào
-Hẻm : Loại lối vào hẻm
-LotShape : Hình dạng chung của bất động sản

-LandContour : Độ phẳng của bất động sản
-Tiện ích : Loại tiện ích có sẵn
-LotConfig : Cấu hình lô
-LandSlope : Độ dốc của bất động sản
-Khu vực lân cận : Vị trí thực tế trong giới hạn thành phố Ames
-Điều kiện 1 : Gần đường chính hoặc đường sắt
-Điều kiện 2 : Gần đường chính hoặc đường sắt (nếu có điều kiện thứ hai)
-BldgType : Loại nhà ở
-HouseStyle : Phong cách nhà ở
-OverallQual : Chất lượng vật liệu và hoàn thiện tổng thể
-OverallCond : Đánh giá tình trạng chung
-Năm xây dựng : Ngày xây dựng ban đầu
-YearRemodAdd : Ngày cải tạo
-RoofStyle : Kiểu mái nhà
-RoofMatl : Vật liệu lợp mái
-Exterior1st : Lớp phủ bên ngoài ngôi nhà
-Exterior2nd : Lớp phủ bên ngoài ngôi nhà (nếu có nhiều hơn một vật liệu)
-MasVnrType : Loại ván ép xây
-MasVnrArea : Diện tích lớp ốp tường tính bằng feet vuông
-ExterQual : Chất lượng vật liệu ngoại thất
-ExterCond : Tình trạng hiện tại của vật liệu ở bên ngoài
-Nền móng : Loại nền móng
-BsmtQual : Chiều cao tầng hầm
-BsmtCond : Tình trạng chung của tầng hầm
-BsmtExposure : Tường tầng hầm có thể đi ra ngoài hoặc có vườn
-BsmtFinType1 : Chất lượng khu vực hoàn thiện tầng hầm
-BsmtFinSF1 : Loại 1 hoàn thiện feet vuông
-BsmtFinType2 : Chất lượng của khu vực hoàn thiện thứ hai (nếu có)
-BsmtFinSF2 : Loại 2 hoàn thiện feet vuông
-BsmtUnfSF : Diện tích tầng hầm chưa hoàn thiện tính bằng feet vuông
-TotalBsmtSF : Tổng diện tích tầng hầm
-Sưởi ấm : Loại sưởi ấm
-HeatingQC : Chất lượng và tình trạng sưởi ấm
-CentralAir : Điều hòa không khí trung tâm
-Điện : Hệ thống điện
-Tầng 1 : Diện tích sàn tầng 1
-Tầng 2 SF : Diện tích tầng 2
-LowQualFinSF : Diện tích hoàn thiện chất lượng thấp (tất cả các tầng)
-GrLivArea : Diện tích sinh hoạt trên mặt đất (feet vuông)
-BsmtFullBath : Phòng tắm đầy đủ ở tầng hầm

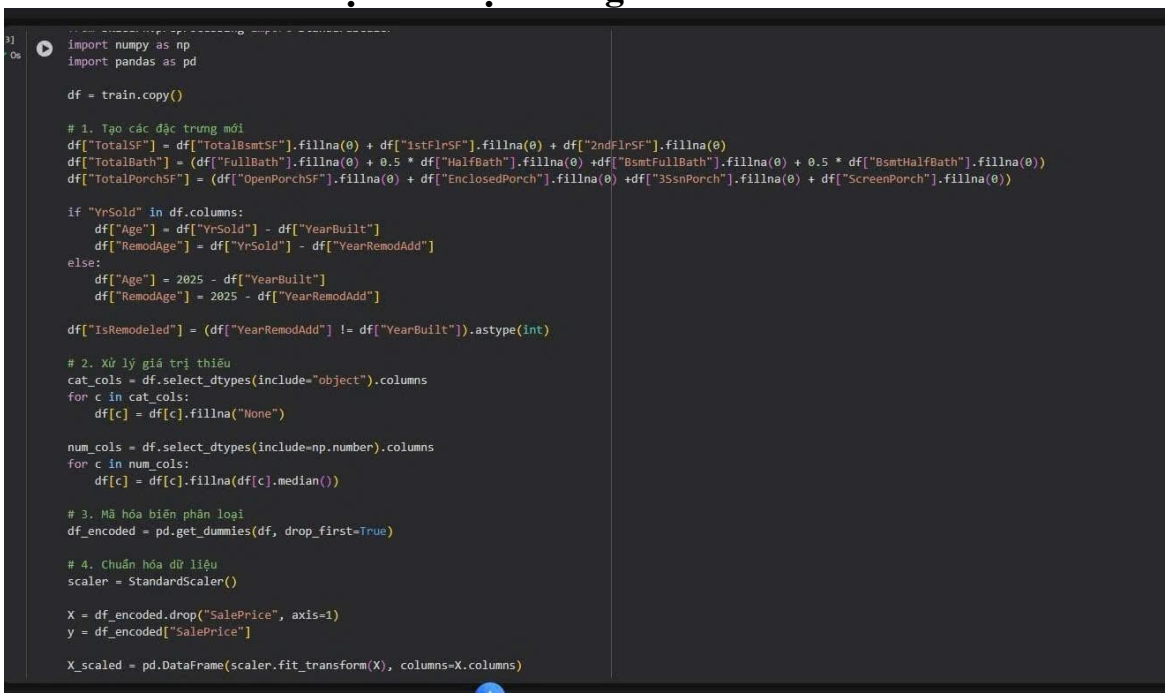
-BsmtHalfBath : Phòng tắm nửa tầng hầm
-FullBath : Phòng tắm đầy đủ trên tầng cao
-HalfBath : Phòng tắm nửa trên mặt đất
-Phòng ngủ : Số phòng ngủ ở trên tầng hầm
-Bếp : Số lượng bếp
-KitchenQual : Chất lượng nhà bếp
-TotRmsAbvGrd : Tổng số phòng trên mặt đất (không bao gồm phòng tắm)
-Chức năng : Đánh giá chức năng của ngôi nhà
-Lò sưởi : Số lượng lò sưởi
-FireplaceQu : Chất lượng lò sưởi
-GarageType : Vị trí gara
-GarageYrBlt : Năm xây dựng nhà để xe
-GarageFinish : Hoàn thiện nội thất của gara
-GarageCars : Kích thước của gara tính theo sức chứa ô tô
-GarageArea : Diện tích của gara tính bằng feet vuông
-GarageQual : Chất lượng gara
-GarageCond : Tình trạng nhà để xe
-PavedDrive : Đường lái xe lát đá
-WoodDeckSF : Diện tích sàn gỗ tính bằng feet vuông
-OpenPorchSF : Diện tích hiên mở tính bằng feet vuông
-EnclosedPorch : Diện tích hiên kín tính bằng feet vuông
-3SsnPorch : Diện tích hiên ba mùa tính bằng feet vuông
-ScreenPorch : Diện tích hiên có màn che tính bằng feet vuông
-PoolArea : Diện tích hồ bơi tính bằng feet vuông
-PoolQC : Chất lượng hồ bơi
-Hàng rào : Chất lượng hàng rào
-MiscFeature : Tính năng khác không có trong các danh mục khác
-MiscVal : Giá trị của tính năng khác nhau
-MoSold : Tháng bán ra
-YrSold : Năm bán
-SaleType : Loại hình bán hàng
-SaleCondition : Tình trạng bán hàng

2.2.1. Tải dữ liệu từ máy lên:



```
!pip install lightgbm xgboost scikit-learn matplotlib seaborn joblib --qu

# 1 Import thư viện
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split

# 2 Hiển thị full cột
pd.set_option('display.max_columns', None)

# 3 Đọc dữ liệu
train = pd.read_csv("train.csv")
test = pd.read_csv("test.csv")
sample = pd.read_csv("sample_submission.csv")

print("Dữ liệu train:", train.shape)
print(" Dữ liệu test:", test.shape)

# 4 Xem 5 dòng đầu
display(train.head(5))

# 5 Kiểm tra cột đích
```

2.2.2. Tiền xử lí và tạo các đặc trưng:



```
import numpy as np
import pandas as pd

df = train.copy()

# 1. Tạo các đặc trưng mới
df["TotalSF"] = df["TotalBsmtSF"].fillna(0) + df["1stFlrSF"].fillna(0) + df["2ndFlrSF"].fillna(0)
df["TotalBath"] = (df["FullBath"].fillna(0) + 0.5 * df["HalfBath"].fillna(0) +df["BsmtFullBath"].fillna(0) + 0.5 * df["BsmtHalfBath"].fillna(0))
df["TotalPorchSF"] = (df["OpenPorchSF"].fillna(0) + df["EnclosedPorch"].fillna(0) +df["3SsnPorch"].fillna(0) + df["ScreenPorch"].fillna(0))

if "YrSold" in df.columns:
    df["Age"] = df["YrSold"] - df["YearBuilt"]
    df["RemodAge"] = df["YrSold"] - df["YearRemodAdd"]
else:
    df["Age"] = 2025 - df["YearBuilt"]
    df["RemodAge"] = 2025 - df["YearRemodAdd"]

df["IsRemodeled"] = (df["YearRemodAdd"] != df["YearBuilt"]).astype(int)

# 2. Xử lý giá trị thiếu
cat_cols = df.select_dtypes(include="object").columns
for c in cat_cols:
    df[c] = df[c].fillna("None")

num_cols = df.select_dtypes(include=np.number).columns
for c in num_cols:
    df[c] = df[c].fillna(df[c].median())

# 3. Mã hóa biến phân loại
df_encoded = pd.get_dummies(df, drop_first=True)

# 4. Chuẩn hóa dữ liệu
scaler = StandardScaler()

X = df_encoded.drop("SalePrice", axis=1)
y = df_encoded["SalePrice"]

X_scaled = pd.DataFrame(scaler.fit_transform(X), columns=X.columns)
```

2.2.3. Heat map:

	SalePrice
SalePrice	1.000000
OverallQual	0.790982
TotalSF	0.782260
GrLivArea	0.708624
GarageCars	0.640409
TotalBath	0.631731
GarageArea	0.623431
TotalBsmtSF	0.613581
1stFlrSF	0.605852
FullBath	0.560664
TotRmsAbvGrd	0.533723
YearBuilt	0.522897
YearRemodAdd	0.507101
Foundation_PConc	0.497734
MasVnrArea	0.472614

dtype: float64

-Tìm ra các dặc trưng ảnh hưởng đến giá nhà:

+OverallQual 0.79: Chất lượng tổng thể của ngôi nhà có ảnh hưởng mạnh nhất đến giá bán.

+TotalSF (biến mới tạo) 0.78 Tổng diện tích sàn (tầng hầm + tầng 1 + tầng 2) có tương quan rất cao, chứng minh đặc trưng mới hữu ích.

+GrLivArea 0.71 Diện tích sinh hoạt trên mặt đất cũng ảnh hưởng mạnh đến giá.

+GarageCars 0.64 Số lượng chỗ đỗ xe trong gara tỷ lệ thuận với giá bán.

+TotalBath (biến mới tạo) 0.63 Tổng số phòng tắm (bao gồm tầng hầm) có tác động tích cực rõ rệt.

+GarageArea 0.62 Diện tích gara cũng phản ánh phần nào quy mô nhà.

+TotalBsmtSF 0.61 Diện tích tầng hầm góp phần vào diện tích tổng thể.

+1stFlrSF 0.61 Diện tích tầng 1 ảnh hưởng đáng kể đến giá.

+FullBath 0.56 Số phòng tắm đầy đủ là yếu tố tiện nghi liên quan đến giá.

+TotRmsAbvGrd 0.53 Tổng số phòng trên mặt đất thể hiện quy mô và tiện nghi.

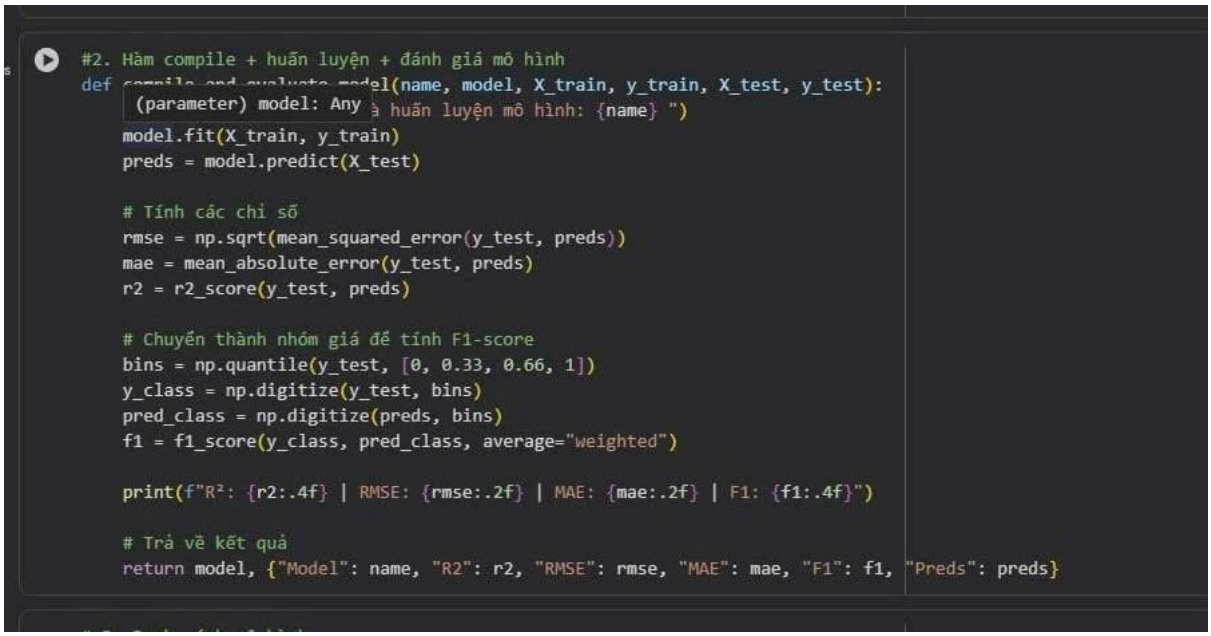
+YearBuilt 0.52 Nhà càng mới thường có giá cao hơn.

+YearRemodAdd 0.51 Nhà được sửa hoặc nâng cấp gần đây có giá trị cao hơn.

+Foundation_PConc 0.50 Loại móng bê tông đúc sẵn (PConc) thường đi kèm nhà chất lượng cao.

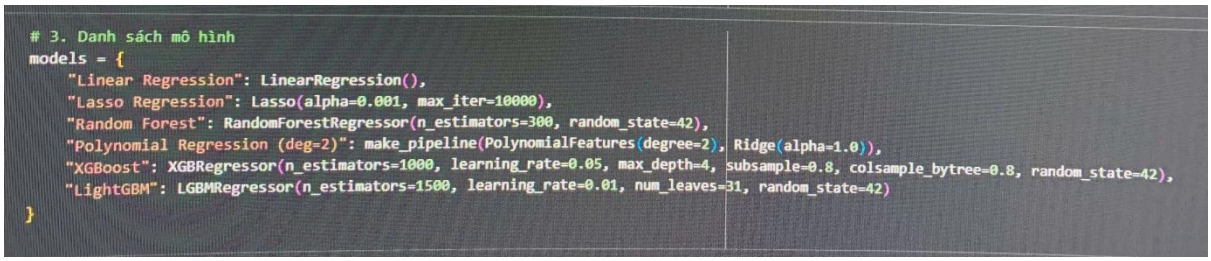
+MasVnrArea 0.47 Diện tích tường ốp đá/gạch có tương quan khá mạnh với giá bán.

2.2.4. Compile, huấn luyện và đánh giá mô hình:



```python
#2. Hàm compile + huấn luyện + đánh giá mô hình
def compile_and_evaluate_model(name, model, X_train, y_train, X_test, y_test):
    (parameter) model: Any à huấn luyện mô hình: {name} ")
    model.fit(X_train, y_train)
    preds = model.predict(X_test)

    # Tính các chỉ số
    rmse = np.sqrt(mean_squared_error(y_test, preds))
    mae = mean_absolute_error(y_test, preds)
    r2 = r2_score(y_test, preds)

    # Chuyển thành nhóm giá để tính F1-score
    bins = np.quantile(y_test, [0, 0.33, 0.66, 1])
    y_class = np.digitize(y_test, bins)
    pred_class = np.digitize(preds, bins)
    f1 = f1_score(y_class, pred_class, average="weighted")

    print(f"R²: {r2:.4f} | RMSE: {rmse:.2f} | MAE: {mae:.2f} | F1: {f1:.4f}")

    # Trả về kết quả
    return model, {"Model": name, "R2": r2, "RMSE": rmse, "MAE": mae, "F1": f1, "Preds": preds}
```

2.2.5. Các mô hình:



```python
# 3. Danh sách mô hình
models = {
    "Linear Regression": LinearRegression(),
    "Lasso Regression": Lasso(alpha=0.001, max_iter=10000),
    "Random Forest": RandomForestRegressor(n_estimators=300, random_state=42),
    "Polynomial Regression (deg=2)": make_pipeline(PolynomialFeatures(degree=2), Ridge(alpha=1.0)),
    "XGBoost": XGBRegressor(n_estimators=1000, learning_rate=0.05, max_depth=4, subsample=0.8, colsample_bytree=0.8, random_state=42),
    "LightGBM": LGBMRegressor(n_estimators=1500, learning_rate=0.01, num_leaves=31, random_state=42)
}
```

-Truyền vào các tham số tương đối để chạy thử mô hình

2.2.6. Kết quả các mô hình:

	Model	R2	RMSE	MAE	F1
4	XGBoost	0.914619	25591.067192	15574.893555	0.855733
5	LightGBM	0.889950	29053.742875	16695.970994	0.871768
2	Random Forest	0.881856	30103.248515	17821.214966	0.849922
3	Polynomial Regression (deg=2)	0.828365	36283.526216	24509.343922	0.686767
1	Lasso Regression	0.411621	67179.327419	22671.409140	0.817222
0	Linear Regression	0.103489	82924.878932	24052.498302	0.817222

-Kết luận:

-Linear Regression và Lasso Regression:

+Cho kết quả R^2 rất thấp (0.10–0.41) → mô hình tuyến tính không thể nắm bắt được mối quan hệ phi tuyến phức tạp trong dữ liệu giá nhà.

+RMSE cao (trên 67.000), chứng tỏ dự đoán kém chính xác.

+Dù F1-score cao, nhưng đây là do phân nhóm tương đối đều, không phản ánh được chất lượng hồi quy.

-Polynomial Regression (bậc 2):

+Cải thiện đáng kể so với hồi quy tuyến tính ($R^2 = 0.83$).

+Tuy nhiên RMSE vẫn còn cao (≈ 36.000), và mô hình có xu hướng overfitting nhẹ, do thêm quá nhiều biến tương tác.

-Random Forest:

+Mô hình cây rừng cho kết quả ổn định, khá chính xác ($R^2 \approx 0.88$, RMSE ≈ 30.000).

+Ưu điểm: dễ huấn luyện, ít cần tinh chỉnh.

+Nhược điểm: dự đoán hơi chậm khi dữ liệu lớn, chưa tối ưu ở nhóm giá cao.

-LightGBM:

+Là mô hình boosting mạnh, đạt $R^2 = 0.89$ và RMSE ≈ 29.000, tốt hơn Random Forest.

+Ưu điểm: chạy nhanh, hiệu suất cao, F1-score cao nhất (0.872).

+Nhược điểm: cần tuning cẩn thận để tránh overfitting.

-XGBoost:

+Là mô hình tốt nhất trong toàn bộ thử nghiệm:

+R^2 = 0.915, RMSE = 25,591, MAE = 15,575.

+Mô hình thể hiện độ khớp rất cao giữa giá dự đoán và giá thực tế, sai số trung bình nhỏ hơn 10% giá nhà thật.

+F1-score (0.856) cũng ở mức tốt, cho thấy mô hình phân loại nhóm giá ổn định.

+Mô hình tổng quát hóa tốt, không bị lệch ở vùng giá thấp, chỉ sai khác nhẹ với nhà giá trị cao.

2.2.7. Tunning với 40 tổ hợp cho kết quả rất tốt:

```
#tuning mô hình với 40 đặc trưng

from sklearn.model_selection import RandomizedSearchCV
from xgboost import XGBRegressor

xgb_model = XGBRegressor(
    objective='reg:squarederror',
    random_state=42,
    n_jobs=-1,
    tree_method='hist'
)

param_dist = {
    "n_estimators": [500, 800, 1000],
    "learning_rate": [0.01, 0.03, 0.05],
    "max_depth": [3, 4, 5, 6],
    "min_child_weight": [1, 3, 5],
    "subsample": [0.6, 0.8, 1.0],
    "colsample_bytree": [0.6, 0.8, 1.0],
    "gamma": [0, 0.1, 0.2]
}

rs = RandomizedSearchCV(
    estimator=xgb_model,
    param_distributions=param_dist,
    n_iter=40,
    scoring="neg_root_mean_squared_error",
    cv=5,
    verbose=2,
    n_jobs=-1,
    random_state=42
)

rs.fit(X_scaled, np.log1p(y))

print("Best parameters:", rs.best_params_)
print("Best CV RMSE :", -rs.best_score_)
```

-Kết quả:

+R^2: 0.9862

+RMSE: 10,293.49

+MAE: 6,987.03

+F1-score: 0.9082

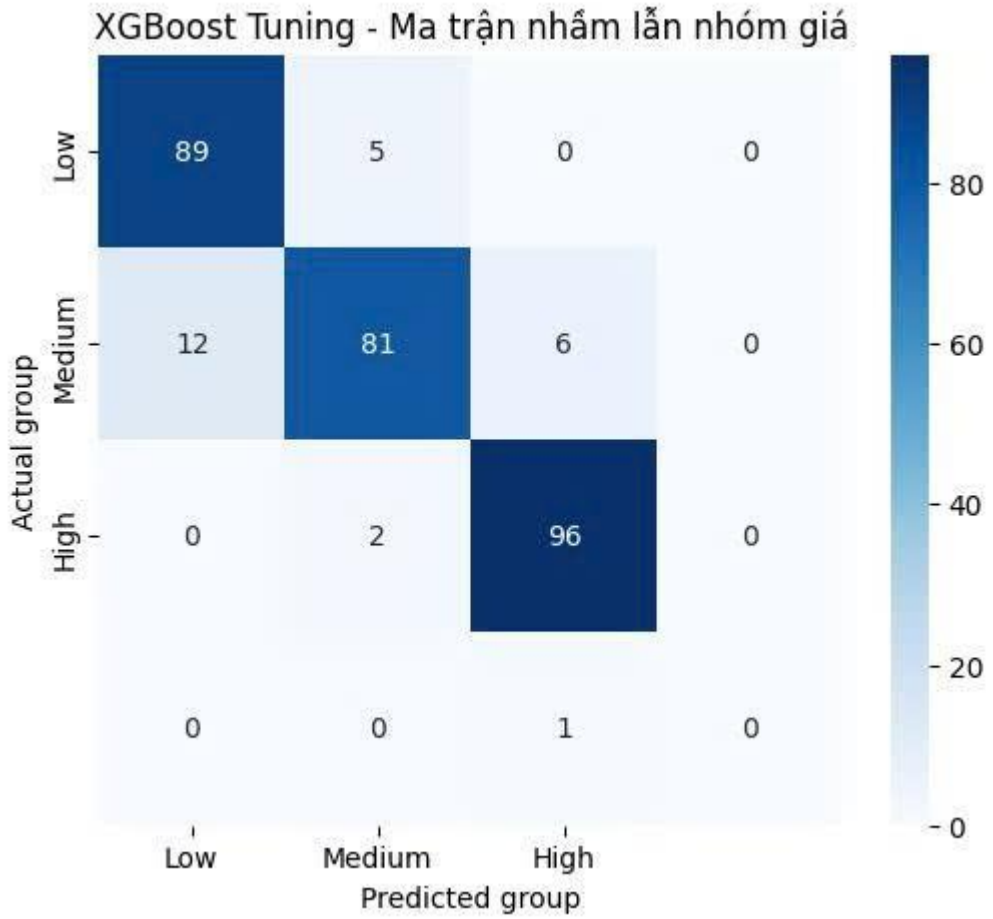
-Nhận xét:

Nhận xét: $R^2 = 0.9862$ → Mô hình giải thích khoảng 98.6% phương sai của giá nhà. Đây là giá trị rất cao, cho thấy mô hình fit dữ liệu rất tốt.

RMSE = 10,293.49 → Sai số trung bình theo chuẩn bình phương khoảng 10.3 nghìn (giả sử đơn vị là USD). So với giá nhà trung bình, đây là mức sai số khá thấp.

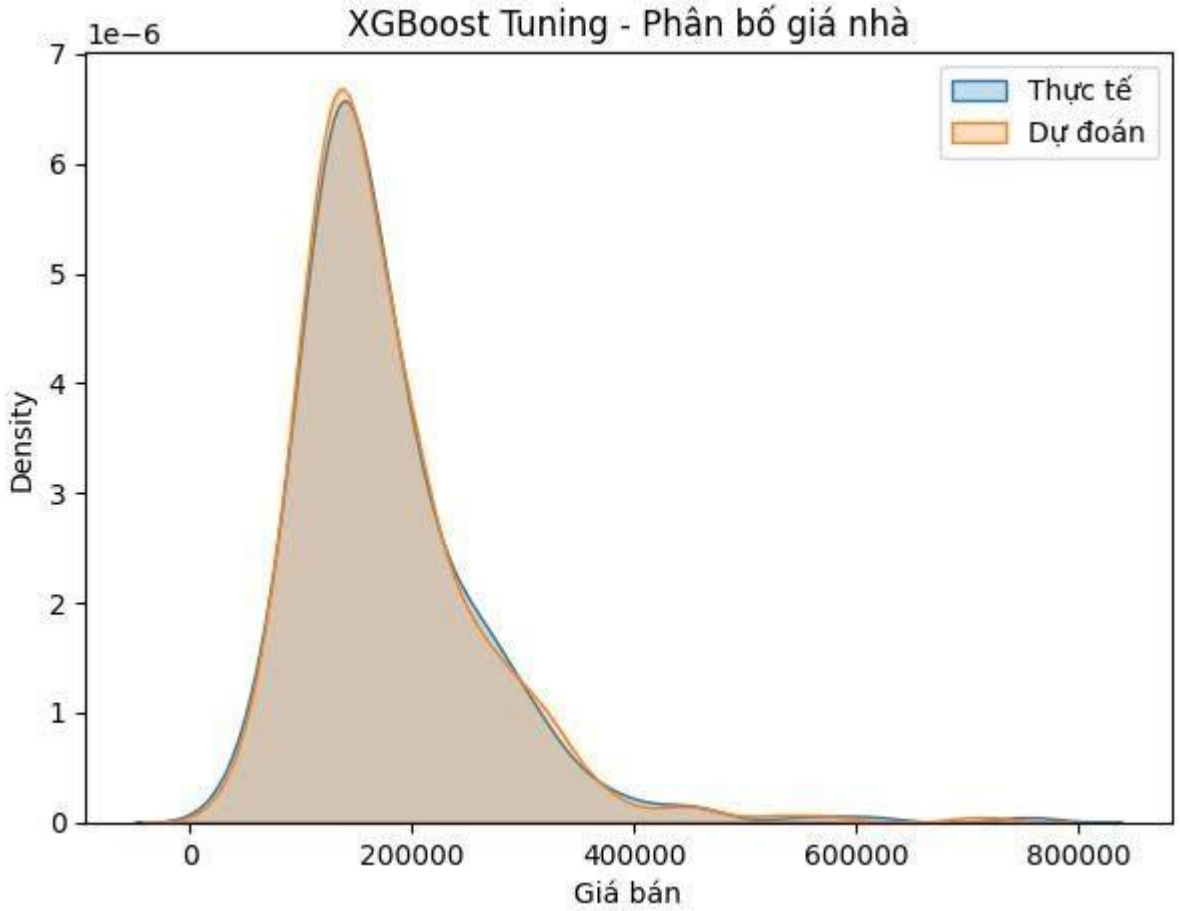
MAE = 6,987.03 → Sai số tuyệt đối trung bình khoảng 7 nghìn, nghĩa là dự đoán trung bình lệch khoảng 7 nghìn so với giá thật, cũng khá hợp lý.

2.2.8. Ma trận nhầm lẫn:



Nhận xét: Mô hình này rất tốt chỉ nhầm lẫn 1 chút ở trung bình

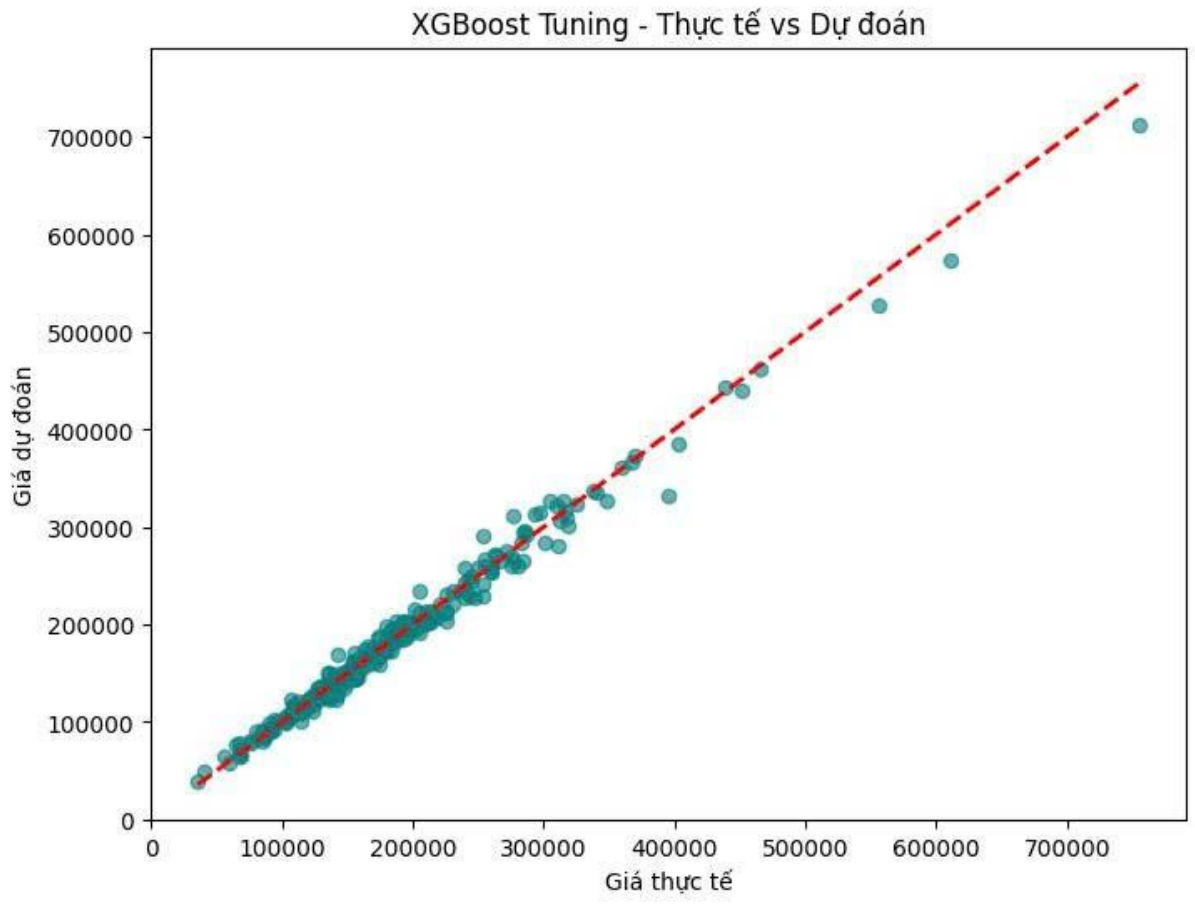
2.2.9. KDE plot so sánh phân bố giá nhà thực tế:



-KDE plot cho thấy mô hình dự đoán phân bố giá tổng thể rất tốt.

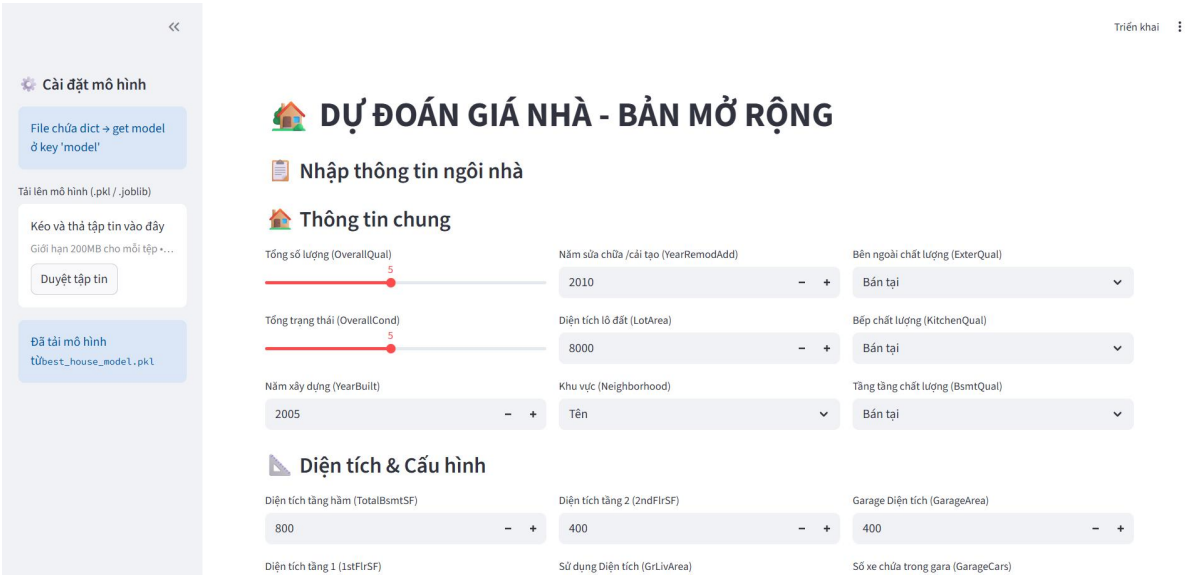
-Kết hợp với $R^2 = 0.9862$, RMSE ≈ 10,293, MAE ≈ 6,987 → model gần như khớp hoàn hảo về tổng thể.

2.2.10. Thực tế với dự đoán:



Nhận xét: Best model giá cực cao có thể lệch 1 chút nhưng tổng thể good.

2.2.11. giao diện web:



-Người dùng nhập các thông số đặc trưng cho căn nhà.

-Ứng dụng gửi dữ liệu đến mô hình đã được huấn luyện.

-Mô hình xử lý và trả về giá trị dự đoán (SalePrice) – tức là giá bán ước tính của căn nhà.

-Kết quả hiển thị trực tiếp trên giao diện.

CHƯƠNG 3 _ KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

3.1. Kết luận:

-Mô hình đã tunnig xgboost đã cho kết quả và sai số tốt nhất với sai số rmse chênh lệch khoảng 5000 đô ở những ngôi nhà cao cấp và điểm f1 là 0.95 suy ra điều này cho thấy mô hình này dự đoán rất mạnh và chính xác, đáng tin cậy

3.2. Hướng phát triển:

- Trang web sẽ tích hợp google maps tại khu vực mà dự đoán

- Áp dụng mô hình đã huấn luyện cho các dự án dự đoán bất động sản khác

- Tích hợp mô hình lên những trang web hiện đại

CHƯƠNG 4_ TÀI LIỆU THAM KHẢO

-Kaggle, youtube

-https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques