TRƯỜNG CÔNG NGHỆ THÔNG TIN

ĐẠI HỌC PHENIKAA

BÁO CÁO BÀI TẬP LỚN

PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN GIÁ NHÀ Ở "AMES IOWA" HOA KỲ

Lớp : Lập trình phân tích dữ liệu với Python-2-2-24(N01)

Giáo viên hướng dẫn: Th.S Nguyễn Văn Thiệu

Sinh viên thực hiện: Hoàng Thị Khuyên 21010588

Nhóm thực hiện: Nhóm 9

HÀ NỘI, THÁNG 5 NĂM 2025

TRƯỜNG CÔNG NGHỆ THÔNG TIN

ĐẠI HỌC PHENIKAA

BÁO CÁO BÀI TẬP LỚN

PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN GIÁ NHÀ Ở "AMES IOWA" HOA KỲ

Lớp : Lập trình phân tích dữ liệu với Python-2-2-24(N01)

Giáo viên hướng dẫn: Th.S Nguyễn Văn Thiệu

Sinh viên thực hiện: Hoàng Thị Khuyên 21010588

Nhóm thực hiện: Nhóm 9

HÀ NỘI, THÁNG 5 NĂM 2025

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

BẢNG PHÂN CÔNG CÔNG VIỆC

Họ và tên	Công việc
Hoàng Thị Khuyên	Thu thập dữ liệu, tiền xử lý dữ liệu, phân tích dữ liệu và khám pháp (EDA), lựa chọn đặc trưng, xây dựng mô hình, nhóm ưu mô hình, viết báo cáo và trình bày kết quả

MỞ ĐẦU

Trong bối cảnh kinh tế hiện đại, thị trường bất động sản giữ vai trò then chốt trong sự phát triển chung của nền kinh tế quốc gia. Tại Hòa Kỳ - một trong những quốc gia có nền kinh tế phát triển hàng đầu thế giới - nhu cầu về nhà ở luôn là một chủ đề nóng. Sự bùng nổ dân số đô thị, tốc độ đô thị hóa nhanh chóng cùng với những biến động sau đại dịch Covid 19 đã khiến giá nhà tăng mạnh, kéo theo sự mất cân đối giữa cung và cầu.

Giá nhà ở không chỉ phụ thuộc vào một yếu tố cố định mà là kết quả tổng hợp của nhiều yếu tố khác nhau như diện tích, số lượng phòng, số lượng tầng, … Do đó việc phân tích và sử dụng các mô hình học máy để tìm hiểu và đánh giá mức độ ảnh hưởng của các yếu tố đó là cần thiết và có ý nghĩa thực tiễn cao.

Trong dự án này, thông qua việc thu thập dữ liệu từ Kaggle, áp dụng các thư viện và công cụ phân tích dữ liệu trong Python để tiền xử lý, trực quan hóa và xây dựng mô hình đánh giá. Kết quả kỳ vọng sẽ giúp nhận diện được các yếu tố có tác động lớn đến giá nhà, từ đó đưa ra những gợi ý hữu ích cho người mua,

NỘI DUNG

I. Mục đích nghiên cứu.

Mục đích chính của dự án này là phân tích và xác định những yếu tố quan trọng nhất ảnh hưởng đến giá nhà tại "ames iowa" Hoa Kỳ thông qua việc xử lý và mô hình hóa dữ liệu.

Quy trình thực hiện bao gồm các bước chính như thu thập dữ liệu, xử lý dữ liệu thiếu và mã hóa biến phân loại, phân tích khám phá dữ liệu (EDA) để hiểu rõ mối quan hệ giữa các biến, và cuối cùng là huấn luyện các mô hình dự đoán như Linear Regression và Random Forest Regressor, Polynomial Linear Regression. Các mô hình được đánh giá dựa trên nhiều chỉ số như MAE, MSE RMSE và hệ số xác định R^2.

II. Phương pháp nghiên cứu.

1. Cơ sở lý thuyết.

1. NumPy (Numerical Python) là một thư viện hỗ trợ xử lý dữ liệu số và tính toán khoa học hiệu quả trong Python. Nó cung cấp các cấu trúc dữ liệu mảng nhiều chiều (ndarray), cùng với các hàm toán học để thực hiện các phép tính đại số tuyến tính, thống kê và xử lý dữ liệu số nhanh chóng. NumPy là nền tảng cho nhiều thư viện phân tích và học máy khác trong Python.

2. Pandas là thư viện chuyên xử lý và phân tích dữ liệu dạng bảng (dữ liệu có hàng và cột). Với hai cấu trúc chính là Series và DataFrame, Pandas cho phép thao tác dữ liệu linh hoạt như đọc/ghi từ tệp CSV, xử lý dữ liệu thiếu, gộp nhóm, lọc, sắp xếp và tổng hợp dữ liệu. Đây là công cụ cốt lõi cho mọi dự án phân tích dữ liệu trong Python.

3. Matplotlib: là một thư viện vẽ đồ thị trong Python. Nó cung cấp các công cụ cho việc tạo ra các biểu đồ, đồ thị, histogram và các loại biểu đồ khác để trình bày dữ liệu một cách trực quan và dễ hiểu.

4. Seaborn là một thư viện trực quan hóa dữ liệu được xây dựng dựa trên Matplotlib, giúp tạo ra các biểu đồ thống kê với thiết kế đẹp mắt và trực quan hơn. Seaborn cung cấp các biểu đồ như heatmap, violin plot, boxplot, pairplot,… và rất hữu ích trong việc khám phá mối quan hệ giữa các biến trong tập dữ liệu.

5. Sklearn (Scikit-learn): là một thư viện trong Python cung cấp một tập các công cụ xử lý các bài toán machine learning và statistical modeling. Nó cung cấp các thuật toán phổ biến như hồi quy tuyến tính (linear regression), phân loại (classification), phân cụm (clustering), và rừng ngẫu nhiên (random forest). Sklearn cũng cung cấp các công cụ để tiền xử lý dữ liệu, chọn đặc trưng (feature selection), và đánh giá mô hình.

2. Thu thập và mô tả dữ liệu.

Dữ liệu được lấy về từ Kaggle: "House Prices - Advanced Regression Techniques | Kaggle".

Dữ liệu được sử dụng trong nghiên cứu này được thu thập từ nền tảng Kaggle, một trang web trực tuyến nổi tiếng chuyên tổ chức các cuộc thi về khoa học dữ liệu, học máy và trí tuệ nhân tạo. Cụ thể, bộ dữ liệu được lấy từ cuộc thi mang tên "House Prices - Advanced Regression Techniques", một trong những cuộc thi phổ biến và được nhiều người tham gia nhất trên nền tảng này. Mục tiêu chính của cuộc thi là xây dựng một mô hình học máy có khả năng dự đoán chính xác giá bán của các căn nhà tại thành phố Ames, thuộc bang Iowa, Hoa Kỳ.

Điểm đặc biệt của bộ dữ liệu này là nó cung cấp một tập hợp rất đa dạng và chi tiết các đặc trưng (features) của mỗi căn nhà, bao gồm từ những yếu tố cơ bản như diện tích, số phòng ngủ, số phòng tắm, đến những đặc điểm nâng cao như chất lượng vật liệu xây dựng, năm xây dựng, điều kiện bảo trì, vị trí địa lý và nhiều thông tin liên quan đến cấu trúc nội thất và ngoại thất. Chính vì vậy, bộ dữ liệu này thường được sử dụng như một bài toán mẫu kinh điển trong lĩnh vực hồi quy nâng cao (advanced regression), rất phù hợp cho việc huấn luyện và đánh giá các mô hình học máy hiện đại, đồng thời giúp người học nắm vững kỹ năng tiền xử lý dữ liệu, lựa chọn đặc trưng và hiệu chỉnh mô hình.

Dữ liệu bao gồm 4 tệp chính:

- train.csv: Chứa thông tin của 1.460 căn nhà đã được bán, với hơn 70 biến đặc trưng mô tả đặc điểm vật lý và tình trạng pháp lý của mỗi căn nhà như diện tích sinh hoạt, kiểu nhà, số tầng, năm xây dựng, chất lượng vật liệu,

diện tích gara,.... Đặc biệt, cột SalePrice là biến mục tiêu thể hiện giá bán của căn nhà.

- test.csv: Tương tự như train.csv, nhưng không bao gồm cột SalePrice. Dữ liệu này được dùng để kiểm tra mô hình dự đoán giá sau khi huấn luyện từ tập train.csv.
- data_description.txt: Là tài liệu mô tả chi tiết ý nghĩa của từng cột trong tập dữ liệu. Tệp này được biên soạn bởi Dean De Cock và đã được chỉnh sửa nhẹ để phù hợp với tên cột hiện tại. Đây là tài liệu tham khảo quan trọng giúp hiểu đúng và đầy đủ về các thuộc tính có trong dữ liệu.
- sample_submission.csv: Là một ví dụ minh họa định dạng chuẩn của kết quả dự đoán khi nộp lên hệ thống Kaggle. Tệp này chứa các ID tương ứng với các căn nhà trong test.csv và một cột SalePrice được dự đoán bằng mô hình hồi quy tuyến tính đơn giản sử dụng một số đặc trưng cơ bản như diện tích lô đất, số phòng ngủ, năm bán,...

Dữ liệu sử dụng để thực thi bài toán

Data	Số sample	Số feature
train.csv	1460	81

Bảng1: Mô tả tổng quát bộ dữ liệu

Các đặc tính liên quan đến tập dữ liệu:

Column	Insight	Non-Null Count	Dtype
Id	Mã nhà	2919 non-null	int64
MSSubClass	Loại hình nhà ở	2919 non-null	int64
MSZoning	Phân vùng quy	2919 non-null	object

	hoạch		
LotFrontage	Chiều dài mặt đường	2433 non-null	float64
LotArea	Diện tích lô đất (sqft)	2919 non-null	object
Street	Loại đường tiếp cận	2919 non-null	object
Alley	Lối hẻm tiếp cận	198 non-null	object
LotShape	Hình dạng lô đất	2919 non-null	object
LanContour	Độ bằng phẳng của đất	2919 non-null	object
Utilities	Tiện tích có sẵn	2917 non-null	object
LotConfig	Cấu hình lô đất	2919 non-null	object
LandSlope	Độ dốc lô đất	2919 non-null	object
Neighborhood	Khu phố	2919 non-null	object
Condition1	Tình trạng xung quanh (1)	2919 non-null	object
Condition2	Tình trạng xung quanh (2)	2919 non-null	object
BldgType	Loại nhà	2919 non-null	object
HouseStyle	Phong cách nhà	2919 non-null	object
OverallQual	Chất lượng tổng thể	2919 non-null	int64
OverallCond	Tình trạng tổng thể	2919 non-null	int64
YearBuilt	Năm xây dựng	2919 non-null	int64
YearRemodAdd	Năm sửa chữa	2919 non-null	int64

RoofStyle	Kiểu mái	2919 non-null	object
RoofMatl	Vật liệu mái	2919 non-null	object
Exterior1st	Vật liệu ngoài (1)	2918 non-null	object
Exterior2nd	Vật liệu ngoài (2)	2918 non-null	object
MasVnrType	Loại tường gạch trang trí	1153 non-null	object
MasVnrArea	Diện tích tường gạch (sqft)	2896 non-null	float64
ExterQual	Chất lượng bên ngoài	2919 non-null	object
ExterCond	Tình trạng bên ngoài	2919 non-null	object
Foundation	Móng nhà	2919 non-null	object
BsmtQual	Chiều cao tầng hầm	2838 non-null	object
BsmtCond	Tình trạng tầng hầm	2837 non-null	object
BsmtExposure	Tầm hầm lộ thiên	2837 non-null	object
BsmtFinType1	Loại tầng hầm hoàn thiện (1)	2840 non-null	object
BsmtFinSF1	Diện tích hoàn thiện (1)	2918 non-null	float64
BsmtFinType2	Loại tầng hầm hoàn thiện (2)	2919 non-null	object
BsmtFinSF2	Diện tích hoàn thiện (2)	2919 non-null	float64
BsmtUnfSF	Diện tích chưa hoàn thiện tầng	2919 non-null	float64

	hầm		
TotalBsmtSF	Tổng diện tích tầng hầm	2918 non-null	float64
Heating	Hệ thống sưởi	2919 non-null	object
HeatingQC	Chất lượng hệ thống sưởi	2919 non-null	object
CentralAir	Điều hòa trung tâm	2919 non-null	object
Electrial	Hệ thống điện	2918 non-null	object
1stFlrSF	Diện tích tầng 1	2919 non-null	int64
2ndFlrSF	Diện tích tầng 2	2919 non-null	int64
LowQualFinSF	Diện tích hoàn thiện kém chất lượng	2919 non-null	int64
GrLivArea	Diện tích sinh hoạt trên mặt đất	2919 non-null	int64
BsmtFullBath	Phòng tắm đầy đủ dưới tầm hầng	2917 non-null	float64
BsmtHalfBath	Phòng tắm nhỏ dưới tầng hầm	2917 non-null	float64
FullBath	Phòng tắm đầy đủ trên mặt đất	2919 non-null	int64
HalfBath	Phòng tắm nhỏ trên mặt đất	2919 non-null	int64
BedroomAbvGr	Phòng ngủ trên mặt đất	2919 non-null	int64
KitchenAbvGr	Phòng bếp	2919 non-null	int64
KitchenQual	Chất lượng bếp	2918 non-null	object

TotRmsAbvGrd	Tổng số phòng trên mặt đất	2919 non-null	int64
Functional	Chức năng nhà	2917 non-null	object
Fireplaces	Số lò sưởi	2919 non-null	int64
FireplaceQu	Chất lượng lò sưởi	1499 non-null	object
GarageType	Vị trí gara	2762 non-null	object
GarageYrBlt	Năm xây gara	2760 non-null	float64
GarageFinish	Hoàn thiện bên trong gara	2760 non-null	object
GarageCars	Sức chứa xe trong gara	2918 non-null	float64
GarageArea	Diện tích gara (sqft)	2918 non-null	float64
GarageQual	Chất lượng gara	2760 non-null	object
GarageCond	Chất lượng gara	2760 non-null	object
PavedDrive	Lối vào lát gạch	2919 non-null	object
WoodDeckSF	Diện tích sàn gỗ	2919 non-null	int64
OpenPorchSF	Diện tích hiên mở	2919 non-null	int64
EnclosedPorch	Diện tích hiên kín	2919 non-null	int64
3SsnPorch	Diện tích hiên 3 mùa	2919 non-null	int64
ScreenPorch	Diện tích hiên lưới	2919 non-null	int64
PoolArea	Diện tích hồ bơi	2919 non-null	int64
PoolQC	Chất lượng hồ bơi	10 non-null	object

Fence	Chất lượng hàng rào	571 non-null	object
MiscFeature	Tính năng phụ khác	105 non-null	object
MiscVal	Giá trị tính năng phụ	2919 non-null	int64
MoSold	Tháng bán	2919 non-null	int64
YrSold	Năm bán	2919 non-null	int64
SaleType	Loại hình bán	2919 non-null	object
SaleCondition	Tình trạng bán	2919 non-null	object
SalePrice	Giá bán	1460 non-null	64

Bảng 2: Các đặc tính liên quan đến tập dữ liệu

Các giá trị thống kê cụ thể cho các cột dạng số trong tập dữ liệu:

	count	mean	std	min	25%	50%	75%	max
MSSubClass	1460	56.90	42.30	20.00	20.00	50.00	70.00	190.00
LotFrontage	1201	70.05	24.28	21.00	59.00	69.00	80.00	313.00
LotArea	1460	10516.83	9981.26	1300.00	7553.50	9478.50	11601.50	215245.00
OverallQual	1460	6.10	1.38	1.00	5.00	6.00	7.00	10.00
OverallCond	1460	5.58	1.11	1.00	5.00	5.00	6.00	9.00
YearBuilt	1460	1971.27	30.20	1872.00	1954.00	1973.00	2000.00	2010.00
YearRemodAd	1460	1984.87	20.65	1950.00	1967.00	1994.00	2004.00	2010.00

d								
MasVnrArea	1452	103.69	181.07	0.00	0.00	0.00	166.00	1600.00
BsmtFinSF1	1460	443.64	456.10	0.00	0.00	383.50	712.25	5644.00
BsmtFinSF2	1460	46.55	161.32	0.00	0.00	0.00	0.00	1474.00
BsmtUnfSF	1460	567.24	441.87	0.00	223.00	477.50	808.00	2336.00
TotalBsmtSF	1460	1057.43	438.71	0.00	795.75	991.50	1298.25	6110.00
1stFlrSF	1460	1162.63	386.59	334.00	882.00	1087.00	1391.25	4692.00
2ndFlrSF	1460	346.99	436.53	0.00	0.00	0.00	728.00	2065.00
LowQualFinSF	1460	5.84	48.62	0.00	0.00	0.00	0.00	572.00
GrLivArea	1460	1515.46	525.48	334.00	1129.50	1464.00	1776.75	5642.00
BsmtFullBath	1460	0.43	0.52	0.00	0.00	0.00	1.00	3.00
BsmtHalfBath	1460	0.06	0.24	0.00	0.00	0.00	0.00	2.00
FullBath	1460	1.57	0.55	0.00	1.00	2.00	2.00	3.00
HalfBath	1460	0.38	0.50	0.00	0.00	0.00	1.00	2.00
BedroomAbvGr	1460	2.87	0.82	0.00	2.00	3.00	3.00	8.00

KitchenAbvGr	1460	01.05	0.22	0.00	1.00	1.00	1.00	3.00
TotRmsAbvGrd	1460	6.52	1.63	2.00	5.00	6.00	7.00	14.00
Fireplaces	1460	0.61	0.64	0.00	0.00	1.00	1.00	3.00
GarageYrBlt	1379	1978.51	24.69	1900.00	1961.00	1980.00	2002.00	2010.00
GarageCars	1460	1.77	0.75	0.00	1.00	2.00	2.00	4.00
GarageArea	1460	472.98	213.80	0.00	334.50	480.00	576.00	1418.00
WoodDeckSF	1460	94.24	125.34	0.00	0.00	0.00	168.00	857.00
OpenPorchSF	1460	46.66	66.26	0.00	0.00	25.00	68.00	547.00
EnclosedPorch	1460	21.95	61.12	0.00	0.00	0.00	0.00	552.00
3SsnPorch	1460	3.41	29.32	0.00	0.00	0.00	0.00	508.00
ScreenPorch	1460	15.06	55.76	0.00	0.00	0.00	0.00	480.00
PoolArea	1460	2.76	40.18	0.00	0.00	0.00	0.00	738.00
MiscVal	1460	43.49	496.12	0.00	0.00	0.00	0.00	15500.00
MoSold	1460	6.32	2.70	1.00	5.00	6.00	8.00	12.00
YrSold	1460	2007.82	1.33	2006.00	2007.00	2008.00	2009.00	2010.00
SalePrice	1460	180921.20	79442.50	34900.00	129975.00	163000.00	214000.00	755000.00

Bảng 3: Các chỉ số thống kê mô tả các cột có dạng số.

Mô tả thống kê các cột có dạng là Object:

Tên cột	Giá trị	Giải nghĩa giá trị	Số lượng	Tỷ lệ (%)
MSZoning	RL	Residential Low Density	1151	78.84
	RM	Residential Medium Density	218	14.93
	FV	Floating Village Residential	65	4.45
	RH	Residential High Density	16	1.1
	C (all)		10	0.68
Street	Pave	Paved	1454	99.59
	Grvl	Gravel	6	0.41
Alley	(NaN)		1369	93.77
	Grvl	Gravel	50	3.42
	Pave	Paved	41	2.81
LotShape	Reg	Regular	925	63.36
	IR1	Slightly irregular	484	33.15
	IR2	Moderately Irregular	41	2.81
	IR3	Irregular	10	0.68
LandContour	Lvl	Near Flat/Level	1311	89.79
	Bnk	Banked - Quick and significant rise from street grade to building	63	4.32
	HLS	Hillside - Significant slope from side to side	50	3.42
	Low	Depression	36	2.47

Utilities	AllPub	All public Utilities (E,G,W,& S)	1459	99.93
	NoSeWa	Electricity and Gas Only	1	0.07
LotConfig	Inside	Inside lot	1052	72.05
	Corner	Corner lot	263	18.01
	CulDSac	Cul-de-sac	94	6.44
	FR2	Frontage on 2 sides of property	47	3.22
	FR3	Frontage on 3 sides of property	4	0.27
LandSlope	Gtl	Gentle slope	1382	94.66
	Mod	Moderate Slope	65	4.45
	Sev	Severe Slope	13	0.89
Neighborhood	NAmes		225	15.41
	CollgCr	College Creek	150	10.27
	OldTown	Old Town	113	7.74
	Edwards	Edwards	100	6.85
	Somerst	Somerset	86	5.89
	Gilbert	Gilbert	79	5.41
	NridgHt	Northridge Heights	77	5.27
	Sawyer	Sawyer	74	05.07
	NWAmes	Northwest Ames	73	5.0
	SawyerW	Sawyer West	59	04.04
	BrkSide	Brookside	58	3.97
	Crawfor	Crawford	51	3.49
	Mitchel	Mitchell	49	3.36
	NoRidge	Northridge	41	2.81

	Timber	Timberland	38	2.6
	IDOTRR	Iowa DOT and Rail Road	37	2.53
	ClearCr	Clear Creek	28	1.92
	StoneBr	Stone Brook	25	1.71
	SWISU	South & West of Iowa State University	25	1.71
	MeadowV	Meadow Village	17	1.16
	Blmngtn	Bloomington Heights	17	1.16
	BrDale	Briardale	16	1.1
	Veenker	Veenker	11	0.75
	NPkVill	Northpark Villa	9	0.62
	Blueste	Bluestem	2	0.14
Condition1	Norm	Normal	1260	86.3
	Feedr	Adjacent to feeder street	81	5.55
	Artery	Adjacent to arterial street	48	3.29
	RRAn	Adjacent to North-South Railroad	26	1.78
	PosN	Near positive off-site feature--park, greenbelt, etc.	19	1.3
	RRAe	Adjacent to East-West Railroad	11	0.75
	PosA	Adjacent to postive off-site feature	8	0.55
	RRNn	Within 200' of North-South Railroad	5	0.34

	RRNe	Within 200' of East-West Railroad	2	0.14
Condition2	Norm	Normal	1445	98.97
	Feedr	Adjacent to feeder street	6	0.41
	Artery	Adjacent to arterial street	2	0.14
	RRNn	Within 200' of North-South Railroad	2	0.14
	PosN	Near positive off-site feature--park, greenbelt, etc.	2	0.14
	PosA	Adjacent to postive off-site feature	1	0.07
	RRAn	Adjacent to North-South Railroad	1	0.07
	RRAe	Adjacent to East-West Railroad	1	0.07
BldgType	1Fam	Single-family Detached	1220	83.56
	TwnhsE	Townhouse End Unit	114	7.81
	Duplex	Duplex	52	3.56
	TwnhsI	Townhouse Inside Unit	43	2.95
	2fmCon	Two-family Conversion; originally built as one-family dwelling	31	2.12
HouseStyle	1Story	One story	726	49.73
	2Story	Two story	445	30.48

	1.5Fin	One and one-half story: 2nd level finished	154	10.55
	SLvl	Split Level	65	4.45
	SFoyer	Split Foyer	37	2.53
	1.5Unf	One and one-half story: 2nd level unfinished	14	0.96
	2.5Unf	Two and one-half story: 2nd level unfinished	11	0.75
	2.5Fin	Two and one-half story: 2nd level finished	8	0.55
RoofStyle	Gable	Gable	1141	78.15
	Hip	Hip	286	19.59
	Flat	Flat	13	0.89
	Gambrel	Gabrel (Barn)	11	0.75
	Mansard	Mansard	7	0.48
	Shed	Shed	2	0.14
RoofMatl	CompShg	Standard (Composite) Shingle	1434	98.22
	Tar&Grv	Gravel & Tar	11	0.75
	WdShngl	Wood Shingles	6	0.41
	WdShake	Wood Shakes	5	0.34
	Metal	Metal	1	0.07
	Membran	Membrane	1	0.07

	Roll	Roll	1	0.07
	ClyTile	Clay or Tile	1	0.07
Exterior1st	VinylSd	Vinyl Siding	515	35.27
	HdBoard	Hard Board	222	15.21
	MetalSd	Metal Siding	220	15.07
	Wd Sdng	Wood Siding	206	14.11
	Plywood	Plywood	108	7.4
	CemntBd	Cement Board	61	4.18
	BrkFace	Brick Face	50	3.42
	WdShing	Wood Shingles	26	1.78
	Stucco	Stucco	25	1.71
	AsbShng	Asbestos Shingles	20	1.37
	BrkComm	Brick Common	2	0.14
	Stone	Stone	2	0.14
	AsphShn	Asphalt Shingles	1	0.07
	ImStucc	Imitation Stucco	1	0.07
	CBlock	Cinder Block	1	0.07
Exterior2nd	VinylSd	Vinyl Siding	504	34.52
	MetalSd	Metal Siding	214	14.66
	HdBoard	Hard Board	207	14.18
	Wd Sdng	Wood Siding	197	13.49
	Plywood	Plywood	142	9.73
	CmentBd	Cement Board	60	4.11
	Wd Shng	Wood Siding	38	2.6
	Stucco	Stucco	26	1.78

	BrkFace	Brick Face	25	1.71
	AsbShng	Asbestos Shingles	20	1.37
	ImStucc	Imitation Stucco	10	0.68
	Brk Cmn	Brick Common	7	0.48
	Stone	Stone	5	0.34
	AsphShn	Asphalt Shingles	3	0.21
	Other	Other	1	0.07
	CBlock	Cinder Block	1	0.07
MasVnrType	None	None	864	59.18
	BrkFace	Brick Face	445	30.48
	Stone	Stone	128	8.77
	BrkCmn	Brick Common	15	01.03
	(NaN)		8	0.55
ExterQual	TA	Average/Typical	906	62.05
	Gd	Good	488	33.42
	Ex	Excellent	52	3.56
	Fa	Fair	14	0.96
ExterCond	TA	Average/Typical	1282	87.81
	Gd	Good	146	10.0
	Fa	Fair	28	1.92
	Ex	Excellent	3	0.21
	Po	Poor	1	0.07
Foundation	PConc	Poured Contrete	647	44.32
	CBlock	Cinder Block	634	43.42
	BrkTil	Brick & Tile	146	10.0

	Slab	Slab	24	1.64
	Stone	Stone	6	0.41
	Wood	Wood	3	0.21
	44.45	Typical (80-89 inches)	649	44.45
BsmtQual	Gd	Good (90-99 inches)	618	42.33
	Ex	Excellent (100+ inches)	121	8.29
	(NaN)		37	2.53
	Fa	Fair (70-79 inches)	35	2.4
BsmtCond	TA	Typical - slight dampness allowed	1311	89.79
	Gd	Good	65	4.45
	Fa	Fair - dampness or some cracking or settling	45	03.08
	(NaN)		37	2.53
	Po	Poor - Severe cracking, settling, or wetness	2	0.14
BsmtExposure	No	No Exposure	953	65.27
	Av	Average Exposure (split levels or foyers typically score average or above)	221	15.14
	Gd	Good Exposure	134	9.18
	Mn	Mimimum Exposure	114	7.81
	(NaN)		38	2.6
BsmtFinType1	Unf	Unfinshed	430	29.45
	GLQ	Good Living Quarters	418	28.63
	ALQ	Average Living Quarters	220	15.07

	BLQ	Below Average Living Quarters	148	10.14
	Rec	Average Rec Room	133	9.11
	LwQ	Low Quality	74	05.07
	(NaN)		37	2.53
BsmtFinType2	Unf	Unfinshed	1256	86.03
	Rec	Average Rec Room	54	3.7
	LwQ	Low Quality	46	3.15
	(NaN)		38	2.6
	BLQ	Below Average Living Quarters	33	2.26
	ALQ	Average Living Quarters	19	1.3
	GLQ	Good Living Quarters	14	0.96
Heating	GasA	Gas forced warm air furnace	1428	97.81
	GasW	Gas hot water or steam heat	18	1.23
	Grav	Gravity furnace	7	0.48
	Wall	Wall furnace	4	0.27
	OthW	Hot water or steam heat other than gas	2	0.14
	Floor	Floor Furnace	1	0.07
HeatingQC	Ex	Excellent	741	50.75
	TA	Average/Typical	428	29.32
	Gd	Good	241	16.51
	Fa	Fair	49	3.36
	Po	Poor	1	0.07

CentralAir	Y	Yes	1365	93.49
	N	No	95	6.51
Electrical	SBrkr	Standard Circuit Breakers & Romex	1334	91.37
	FuseA	Fuse Box over 60 AMP and all Romex wiring (Average)	94	6.44
	FuseF	60 AMP Fuse Box and mostly Romex wiring (Fair)	27	1.85
	FuseP	60 AMP Fuse Box and mostly knob & tube wiring (poor)	3	0.21
	Mix	Mixed	1	0.07
	(NaN)		1	0.07
KitchenQual	TA	Typical/Average	735	50.34
	Gd	Good	586	40.14
	Ex	Excellent	100	6.85
	Fa	Fair	39	2.67
Functional	Typ	Typical Functionality	1360	93.15
	Min2	Minor Deductions 2	34	2.33
	Min1	Minor Deductions 1	31	2.12
	Mod	Moderate Deductions	15	01.03
	Maj1	Major Deductions 1	14	0.96
	Maj2	Major Deductions 2	5	0.34
	Sev	Severely Damaged	1	0.07
	(NaN)		690	47.26

	Gd	Good - Masonry Fireplace in main level	380	26.03
FireplaceQu	TA	Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement	313	21.44
	Fa	Fair - Prefabricated Fireplace in basement	33	2.26
	Ex	Excellent - Exceptional Masonry Fireplace	24	1.64
	Po	Poor - Ben Franklin Stove	20	1.37
GarageType	Attchd	Attached to home	870	59.59
	Detchd	Detached from home	387	26.51
	BuiltIn	Built-In (Garage part of house - typically has room above garage)	88	06.03
	(NaN)		81	5.55
	Basment	Basement Garage	19	1.3
	CarPort	Car Port	9	0.62
	2Types	More than one type of garage	6	0.41
GarageFinish	Unf	Unfinished	605	41.44
	RFn	Rough Finished	422	28.9
	Fin	Finished	352	24.11
	(NaN)		81	5.55
GarageQual	TA	Typical/Average	1311	89.79

GarageQual	(NaN)		81	5.55
	Fa	Fair	48	3.29
	Gd	Good	14	0.96
	Ex	Excellent	3	0.21
	Po	Poor	3	0.21
GarageCond	TA	Typical/Average	1326	90.82
	(NaN)		81	5.55
	Fa	Fair	35	2.4
	Gd	Good	9	0.62
	Po	Poor	7	0.48
	Ex	Excellent	2	0.14
PavedDrive	Y	Paved	1340	91.78
	N	Dirt/Gravel	90	6.16
	P	Partial Pavement	30	02.05
PoolQC	(NaN)		1453	99.52
	Gd	Good	3	0.21
	Ex	Excellent	2	0.14
	Fa	Fair	2	0.14
Fence	(NaN)		1179	80.75
	MnPrv	Minimum Privacy	157	10.75
	GdPrv	Good Privacy	59	04.04
	GdWo	Good Wood	54	3.7
	MnWw	Minimum Wood/Wire	11	0.75
MiscFeature	(NaN)		1406	96.3
	Shed	Shed (over 100 SF)	49	3.36

MiscFeature	Gar2	2nd Garage (if not described in garage section)	2	0.14
	Othr	Other	2	0.14
	TenC	Tennis Court	1	0.07
SaleType	WD	Warranty Deed - Conventional	1267	86.78
	New	Home just constructed and sold	122	8.36
	COD	Court Officer Deed/Estate	43	2.95
	ConLD	Contract Low Down	9	0.62
	ConLI	Contract Low Interest	5	0.34
	ConLw	Contract Low Down payment and low interest	5	0.34
	CWD	Warranty Deed - Cash	4	0.27
	Oth	Other	3	0.21
	Con	Contract 15% Down payment regular terms	2	0.14
SaleCondition	Normal	Normal Sale	1198	82.05
	Partial	Home was not completed when last assessed (associated with New Homes)	125	8.56
	Abnorml	Abnormal Sale - trade, foreclosure, short sale	101	6.92
	Family	Sale between family members	20	1.37

	Alloca	Allocation - two linked properties with separate deeds, typically condo with a garage unit	12	0.82
	AdjLand	Adjoining Land Purchase	4	0.27

Bảng 4: Mô tả hống kê các cột có dữ liệu dạng String.

3. Tiền xử lý dữ liệu.

Khi xây dựng mô hình dự đoán giá nhà bằng Linear Regression, một trong những giai đoạn quan trọng nhất là tiền xử lý dữ liệu. Bộ dữ liệu nhà ở từ Kaggle chứa hơn 80 đặc trưng (features) khác nhau, trải rộng từ thông tin số học như diện tích, số tầng, đến các thuộc tính phân loại như khu vực, kiểu nhà, chất lượng vật liệu,... Tuy nhiên, bộ dữ liệu này cũng tồn tại nhiều thách thức trong quá trình xử lý, mà nếu không được giải quyết đúng cách sẽ ảnh hưởng lớn đến độ chính xác và khả năng tổng quát của mô hình.

Một trong những thách thức đầu tiên là dữ liệu bị thiếu. Nhiều cột như Alley, PoolQC, Fence hay FireplaceQu có số lượng giá trị bị thiếu rất lớn, thậm chí hơn 80%. Việc xử lý thiếu dữ liệu không chỉ đơn giản là điền vào chỗ trống, mà còn cần cân nhắc xem đặc trưng đó có thực sự cần thiết không. Nếu không được xử lý cẩn thận, mô hình có thể bị sai lệch hoặc học những mối quan hệ không thực sự tồn tại.

Ngoài ra, nhiều đặc trưng dạng chuỗi (categorical) như Neighborhood, HouseStyle, RoofMatl... cần được mã hóa (encoding) thành dạng số để sử dụng trong Linear Regression. Việc mã hóa không làm mất ý nghĩa phân loại.

3.1 Xử lý các giá trị NaN/ Null.

Bộ dữ liệu có nhiều ô bị thiếu giá trị. Điều này làm giảm chất lượng mô hình nếu không được xử lý đúng. Để giải quyết, nhóm sử dụng một hàm tự định nghĩa có tên handle_missing_values(), với cơ chế như sau:

Với các cột số (int64, float64) là: ‘MSSubClass’, ‘LotFrontage’, ‘LotArea’, ‘OverallQual’, ‘OverallCond’, ‘YearBuilt’, ‘YearRemodAdd’, ‘MasVnrArea’, ‘BsmtFinSF1’, ‘'BsmtFinSF2’, ‘BsmtUnfSF’, ‘TotalBsmtSF’, ‘1stFlrSF’, ‘2ndFlrSF’, ‘LowQualFinSF’, ‘GrLivArea’, ‘BsmtFullBath’, ‘BsmtHalfBath’, ‘FullBath’, ‘HalfBath’, ‘BedroomAbvGr’, ‘KitchenAbvGr’, ‘TotRmsAbvGrd’, ‘Fireplaces’, ‘GarageYrBlt’, ‘GarageCars’, ‘GarageArea’, ‘WoodDeckSF’, ‘OpenPorchSF’, ‘EnclosedPorch’, ‘3SsnPorch’, ‘ScreenPorch’, ‘PoolArea’, ‘MiscVal’, ‘MoSold’, ‘YrSol’', ‘SalePrice’.

Thay thế NaN bằng giá trị trung vị (median), vì nó không bị ảnh hưởng bởi giá trị ngoại lệ(outliers), giúp giữ ổn định phân phối dữ liệu.

Giả sử dữ liệu gốc như sau:

LotFrontage	SalePrice
80	200000
NaN	180000
75	195000
NaN	210000

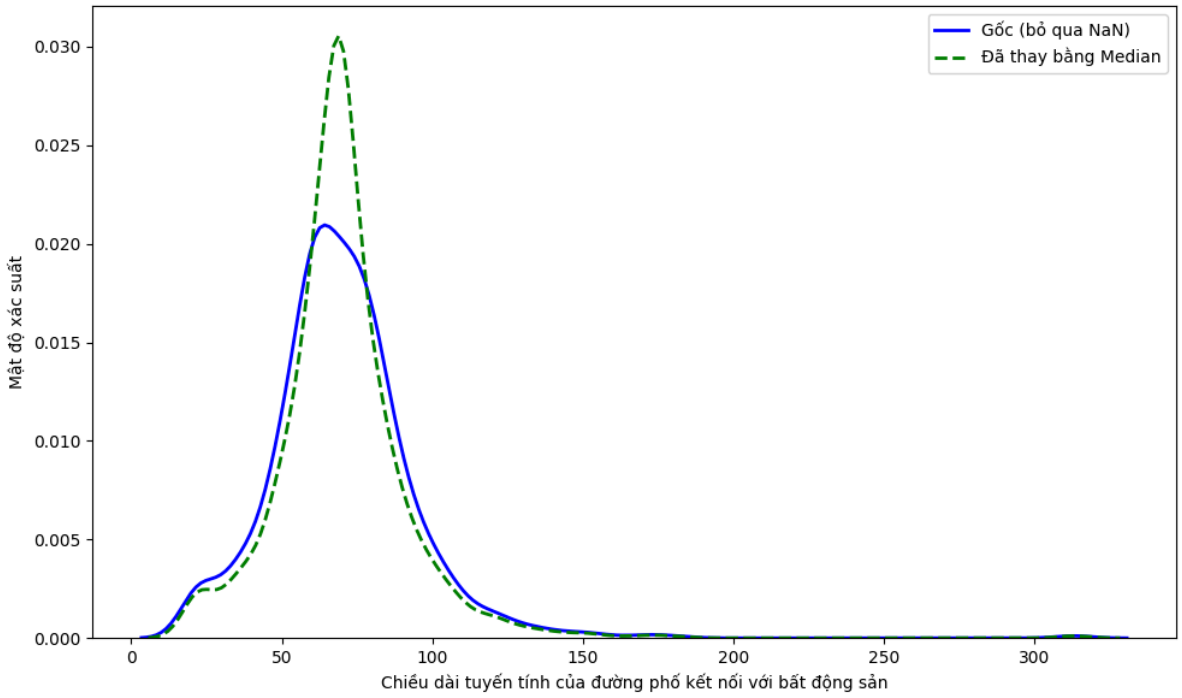
Bảng 5: Mẫu giá trị NaN ở các cột có kiểu dữ liệu dạng số trước khi được xử lý

Sau khi áp dụng hàm:

LotFrontage	SalePrice
80	200000
77.5	180000
75	195000
77.5	210000

Bảng 6: Mẫu giá trị NaN ở các cột có kiểu dữ liệu dạng số sau khi được xử lý.

Nhờ vậy, bộ dữ liệu trở nên đầy đủ và ổn định hơn.



Hinh.1. So sánh ảnh hưởng của việc thay NaN bằng Median đối với phân phối 'LotFronTage'

Với các cột phân loại (object, category): 'MSZoning', 'Street', 'Alley', 'LotShape', 'LandContour', 'Utilities', 'LotConfig', 'LandSlope', 'Neighborhood', 'Condition1', 'Condition2', 'BldgType', 'HouseStyle', 'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'ExterQual', 'ExterCond', 'Foundation',

'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2', 'Heating', 'HeatingQC', 'CentralAir', 'Electrical', 'KitchenQual', 'Functional', 'FireplaceQu', 'GarageType', 'GarageFinish', 'GarageQual', 'GarageCond', 'PavedDrive', 'PoolQC', 'Fence', 'MiscFeature', 'SaleType', 'SaleCondition'.

Thay thế NaN bằng giá trị phổ biến nhất (mode), để đảm bảo tính nhất quán và hợp lý về mặt ngữ nghĩa.

Giả sử dữ liệu gốc như sau:

GarageType	SalePrice
Attchd	200000
Detachd	180000
NaN	195000
Attchd	210000

Bảng 7: Mẫu giá trị NaN ở các cột có kiểu dữ liệu dạng String trước khi được xử lý.

Sau khi áp dụng hàm:

GarageType	SalePrice
Attchd	200000
Detachd	180000
Attchd	195000
Attchd	210000

Bảng 8: Mẫu giá trị NaN ở các cột có kiểu dữ liệu dạng String sau khi được xử lý

3.2. Label Encoder.

Bộ dữ liệu chứa nhiều cột dạng chuỗi (categorical) như 'MSZoning', 'Street', 'Alley', 'LotShape', 'LandContour', 'Utilities', 'LotConfig', 'LandSlope', 'Neighborhood', 'Condition1', 'Condition2', 'BldgType', 'HouseStyle', 'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'ExterQual', 'ExterCond', 'Foundation', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2', 'Heating', 'HeatingQC', 'CentralAir', 'Electrical', 'KitchenQual', 'Functional', 'FireplaceQu', 'GarageType', 'GarageFinish', 'GarageQual', 'GarageCond', 'PavedDrive', 'PoolQC', 'Fence', 'MiscFeature', 'SaleType', 'SaleCondition', khiến các thuật toán học máy không thể xử lý trực tiếp vì chúng yêu cầu dữ liệu dạng số. Nếu không sử dụng Label Encoding hoặc phương pháp mã hóa khác, mô hình sẽ không hiểu được ý nghĩa và mối quan hệ giữa các giá trị này. Ngoài ra, việc giữ nguyên chuỗi có thể gây lỗi trong quá trình huấn luyện, đặc biệt với các mô hình như hồi quy tuyến tính, cây quyết định hoặc SVM. Label Encoder giúp chuyển đổi các giá trị chuỗi thành số nguyên, giúp mô hình học và phân tích dữ liệu hiệu quả hơn.

Ví dụ, Encoder với cột “Utilities”

Giá trị	Ý nghĩa	Encoder
AllPub	All public Utilities (E, G, W & S)	0
ELO	Electricity only	1
NoSeWa	Electricity and Gas Only	2
NoSewr	Electricity, Gas, and Water (Septic Tank)	3

Bảng 9: Ví dụ khi Encoder với cột “Utilities”

3.3. Bỏ cột Id.

Trong bộ dữ liệu, cột Id chỉ đóng vai trò như một định danh duy nhất cho mỗi bản ghi, tức là mỗi căn nhà sẽ có một mã số riêng biệt. Tuy nhiên, cột này không chứa thông tin mô tả đặc điểm của căn nhà, cũng như không có mối quan hệ nào với giá trị mục tiêu cần dự đoán là SalePrice. Do đó, nó không có giá trị sử dụng trong việc xây dựng mô hình học máy.

Nếu giữ lại cột Id trong quá trình huấn luyện, nó có thể gây ra nhiễu cho mô hình, đặc biệt là đối với các thuật toán nhạy cảm với giá trị số hoặc có xu hướng tìm kiếm mẫu trong dữ liệu. Thêm vào đó, cột Id cũng tăng số chiều của dữ liệu một cách không cần thiết, khiến mô hình phải xử lý thêm một biến vô nghĩa, từ đó làm chậm quá trình huấn luyện và giảm hiệu quả tổng thể của mô hình.

3.4. Chuẩn hóa dữ liệu.

Để đảm bảo các đặc trưng (feature) có cùng độ lớn và tránh hiện tượng một số đặc trưng chiếm ưu thế do đó có giá trị tuyệt đối lớn hơn, dữ liệu đã được chuẩn hóa bằng phương pháp StandardScaler. Phương pháp này đưa các giá trị về phân phối chuẩn với trung bình bằng 0 và độ lệch chuẩn bằng 1. Theo công thức:

$$z = \frac{x - \mu}{\sigma}$$

Trong đó:

- x là giá trị gốc
- μ là giá trị trung bình của cột

- σ là độ lệch chuẩn của cột

Ví dụ, nếu xét một vài giá trị ban đầu của đặc trưng GrLivArea (diện tác tầng trệt sinh hoạt) và OverallQual (chất lượng tổng thể):

GrLivArea	Utilities	YearBuilt
1710	0	2003
1262	0	1976
1786	0	2001

Bảng 10: Mẫu các giá trị trước khi chuẩn hóa.

Sau khi chuẩn hóa:

GrLivArea	Utilities	YearBuilt
0.370	-0.026	1.051
-0.498	-0.026	0.157
0.515	-0.026	0.985

Bảng 11: Mẫu các giá trị sau khi chuẩn hóa

4. Khai phá dữ liệu (Exploratory Data Analysis - EDA).

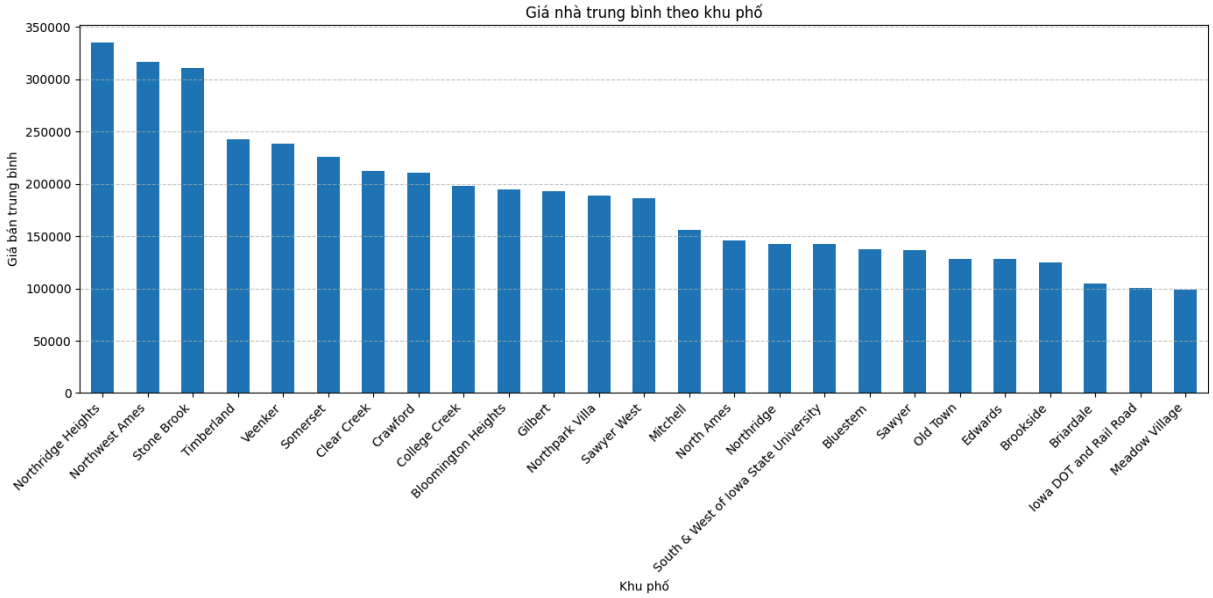
Trong quá trình phân tích dữ liệu, có thể đặt ra một số câu hỏi để khám phá và hiểu rõ thông tin bên trong tập dữ liệu như:

- Nhưng yếu tố tương quan mạnh với giá nhà?
- Giá nhà phân bổ như thế nào ?
- Vật liệu và chất lượng ngôi có ảnh hưởng như thế nào đến giá nhà?

- Giá nhà biến động như thế nào theo từng năm bán ?
- Diện tích ở hợp pháp ảnh hưởng như thế nào đến giá nhà ?

Trong phần này, các biểu đồ sẽ giúp làm sáng tỏ những câu hỏi này.

4.1. Sử dụng barchart thể hiện giá nhà trung bình theo từng khu phố.



Hình 2: Giá nhà theo từng khu phố

Biểu đồ thể hiện giá nhà trung bình theo từng khu phố tại Ames cho thấy sự chênh lệch rõ rệt giữa các khu vực. Những khu phố như NoRidge, NridgHt và StoneBr có mức giá trung bình cao nhất, thường do chất lượng nhà ở vượt trội, vị trí thuận tiện và hạ tầng phát triển. Ngược lại, các khu như MeadowV, IDOTRR hay BrDale có giá thấp hơn đáng kể, có thể do vị trí kém thuận lợi, nhà nhỏ và chất lượng thấp. Điều này phản ánh rõ mối liên hệ giữa vị trí địa lý, điều kiện sống và giá trị bất động sản trong thị trường nhà ở tại Ames.

4.2. Sử dụng heatmap thể hiện mức độ tương quan của các yếu tố với giá nhà.

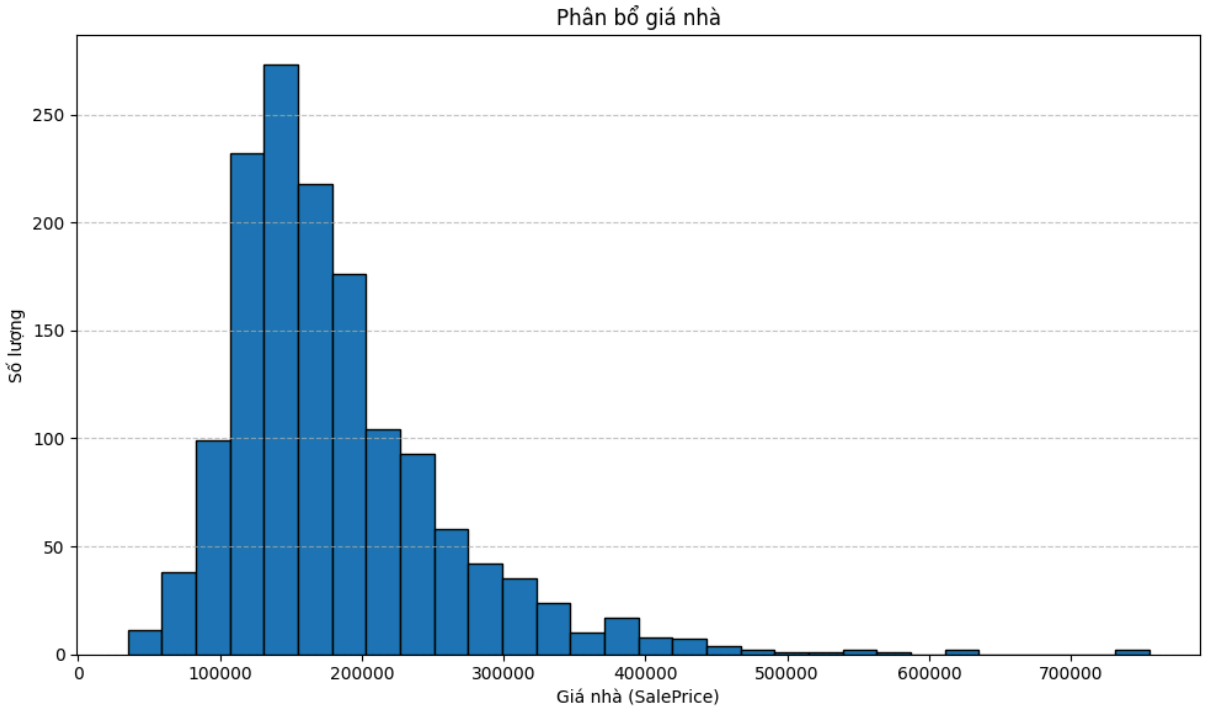
Trong biểu đồ, hai yếu tố liên quan nhất đến giá bán là diện tích sinh hoạt trên mặt đất (tương quan 0.79) và diện tích gara (0.71), cho thấy nhà rộng rãi và có gara lớn thường có giá cao hơn. Ngược lại, tiện ích có sẵn (-0.014) và độ bằng phẳng đất (0.015) hầu như không ảnh hưởng đến giá bán. Kết luận này phản ánh đúng thực tế thị trường bất động sản tại Hoa Kỳ vì:

Diện tích sinh hoạt trên mặt đất và diện tích gara là những yếu tố quan trọng đối với người mua nhà ở "ames iowa". Họ thường ưu tiên không gian sống rộng rãi để phục vụ sinh hoạt gia đình và gara lớn để đỗ xe hoặc làm kho chứa, đặc biệt ở các vùng ngoại ô.

Hình 3: Tương quan giữa các yếu tố và giá nhà

Ngược lại, tiện ích có sẵn như sân chơi hay khu vực chung không phải lúc nào cũng gắn liền với giá trị từng căn nhà cụ thể, vì đó là yếu tố cộng đồng. Độ bằng phẳng của đất cũng ít ảnh hưởng vì ở "ames iowa" nhiều khu dân cư đã được quy hoạch, san nền sẵn, nên yếu tố này không tạo khác biệt đáng kể về giá.

4.3. Sử dụng bar chart trực quan phân bổ dữ liệu của giá nhà.

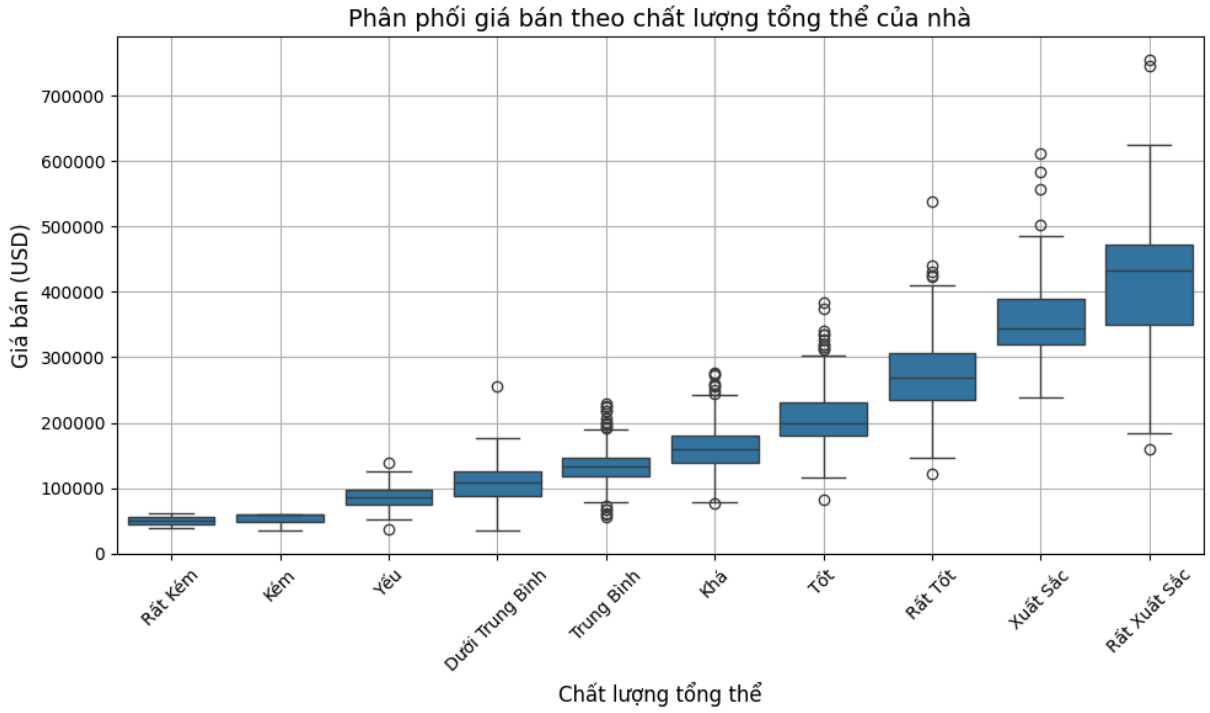


Hình 4. Phân bổ giá nhà

Biểu đồ phân bố giá nhà (SalePrice) cho thấy dữ liệu có dạng phân phối lệch phải, tức là phần lớn các căn nhà trong tập dữ liệu có giá bán nằm ở mức thấp đến trung bình, trong khi chỉ có một số ít căn nhà có giá rất cao. Cụ thể, nhiều căn nhà có mức giá dao động trong khoảng từ 100.000 đến 200.000 đô la Mỹ, và đỉnh phân bố rơi vào khoảng 150.000 đô – đây là mức giá phổ biến nhất trong toàn bộ tập dữ liệu. Ngoài ra, biểu đồ cũng cho thấy sự xuất hiện của một số căn nhà có giá trị cao bất thường, vượt ngưỡng 400.000 đô, thậm chí có trường hợp

đạt nhóm hơn 700.000 đô. Những điểm này xuất hiện với tần suất thấp và có thể được xem là các giá trị ngoại lai (outliers), ảnh hưởng đến phân phối tổng thể của dữ liệu. Bên cạnh đó, các nhà phát triển bất động sản có xu hướng xây dựng nhiều căn nhà ở mức giá này vì dễ bán và phù hợp với thu nhập đại chúng.

4.4. Sử dụng boxplot trực quan phân phối giá bán theo chất lượng tổng thể của của ngôi nhà.

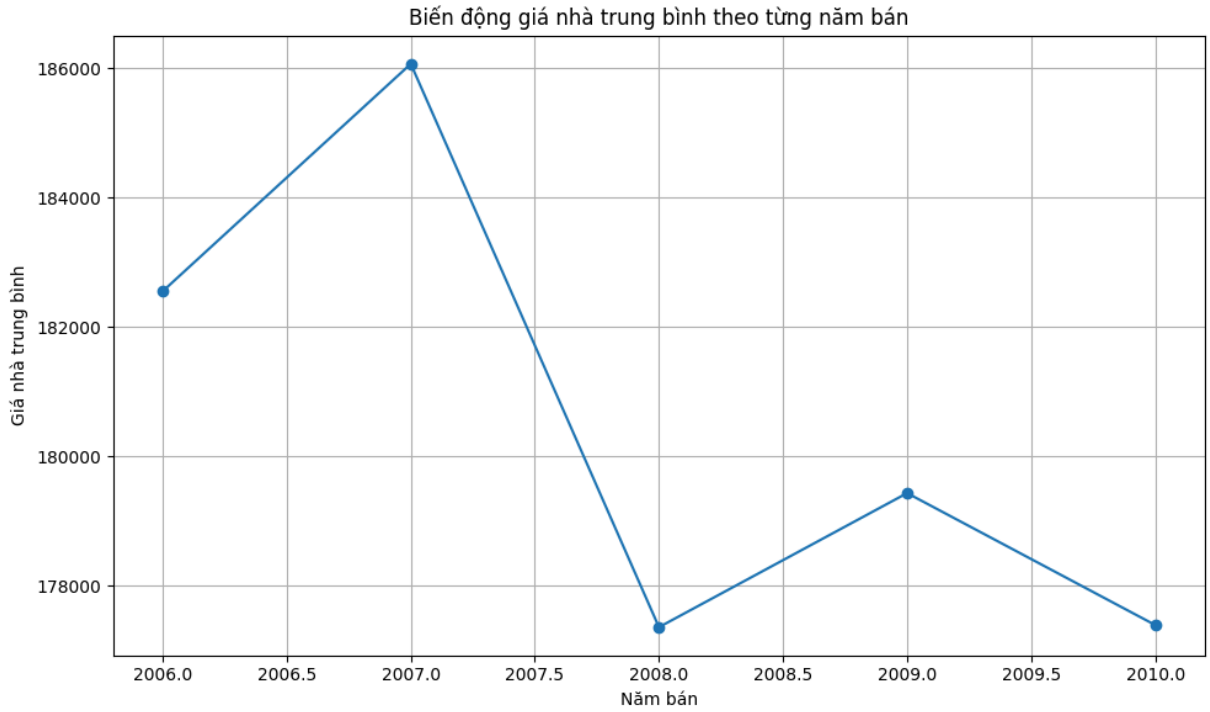


Hình 5. Phân phối giá bán theo chất lượng tổng thể của nhà.

Biểu đồ boxplot thể hiện mối quan hệ giữa chất lượng tổng thể của ngôi nhà và giá bán cho thấy xu hướng rất rõ ràng: khi chất lượng tổng thể tăng lên, giá bán cũng tăng theo một cách đáng kể. Nhóm nhà có chất lượng "Rất kém", "Kém" và "Yếu" có giá bán thấp và tương đối ít dao động, trong khi các nhóm như "Xuất sắc" và "Rất xuất sắc" không chỉ có mức giá trung vị cao hơn nhiều mà còn có khoảng giá dao động rộng hơn, thể hiện sự đa dạng trong các đặc điểm bổ sung của những căn nhà chất lượng cao.

Điều này phản ánh một thực tế phổ biến trong hành vi tiêu dùng bất động sản: người mua sẵn sàng trả giá cao hơn cho những ngôi nhà có chất lượng vượt trội, vì đây không chỉ là nơi ở mà còn là một khoản đầu tư lâu dài. Những căn nhà có chất lượng tốt thường bền vững hơn theo thời gian, ít cần sửa chữa, và giữ được giá trị cao trên thị trường thứ cấp. Hơn nữa, những ngôi nhà này thường đi kèm với vị trí tốt, vật liệu xây dựng chất lượng và thiết kế hiện đại – những yếu tố quan trọng quyết định giá trị tài sản. Vì thế, việc đầu tư vào một căn nhà "Rất tốt" hoặc "Xuất sắc" là một lựa chọn hợp lý đối với nhiều người mua nhằm nhóm ưu giá trị sử dụng lẫn giá trị tài chính trong tương lai.

4.5. Sử dụng biểu đồ đường trực quan giá bán thay đổi theo năm.



Hình 6: Biến động giá nhà trung bình theo từng năm bán

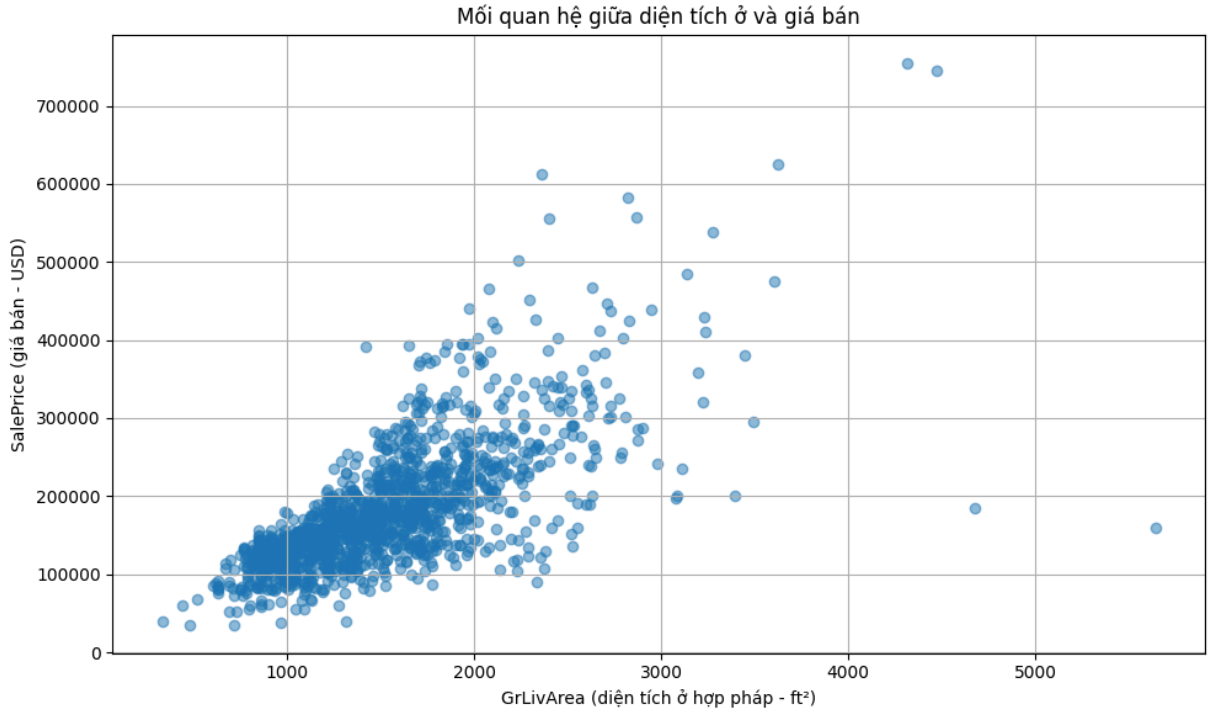
Biểu đồ trên thể hiện sự biến động giá nhà trung bình theo từng năm bán trong giai đoạn từ 2006 đến 2010. Có thể thấy, giá nhà trung bình đạt đỉnh vào năm

2007, với mức cao nhất trong toàn bộ giai đoạn. Đây có thể là thời điểm thị trường bất động sản đang ở giai đoạn sôi động nhất, khi nhu cầu mua nhà cao và niềm tin thị trường vẫn còn mạnh mẽ.

Tuy nhiên, từ năm 2008, giá nhà trung bình giảm mạnh, phản ánh ảnh hưởng rõ rệt của cuộc khủng hoảng tài chính toàn cầu 2007–2008. Tình trạng siết tín dụng, tăng tỷ lệ thất nghiệp và tâm lý e ngại đầu tư khiến giá nhà lao dốc. Dù có sự phục hồi nhẹ trong năm 2009, nhưng giá nhà vẫn không quay lại được mức đỉnh của năm 2007, và tiếp tục suy giảm vào năm 2010.

Điều này cho thấy thị trường bất động sản Mỹ trong giai đoạn này chịu tác động rõ rệt từ yếu tố kinh tế vĩ mô, đặc biệt là khủng hoảng tài chính. Người mua trở nên thận trọng hơn, khả năng vay mượn giảm sút và nhiều ngôi nhà bị bán tháo do mất khả năng thanh toán nợ vay thế chấp, góp phần làm giá nhà giảm sâu.

4.6. Sử dụng scatter plot trực quan mối quan hệ giữa diện tích ở (GrLivArea) và giá bán (SalePrice).



Hình 7. Mối quan hệ giữa diện tích ở và giá bán

Từ biểu đồ scatter, ta thấy có mối tương quan thuận khá rõ ràng giữa diện tích ở hợp pháp (GrLivArea) và giá bán (SalePrice). Nói cách khác, những căn nhà có diện tích sử dụng lớn thường đi kèm với giá bán cao hơn. Tuy nhiên, vẫn tồn tại một số ngoại lệ đáng chú ý. Ví dụ, có những căn nhà có diện tích lớn trên 4000 ft² nhưng lại có giá bán dưới 300000 USD. Điều này cho thấy diện tích không phải là yếu tố duy nhất quyết định giá trị của một ngôi nhà. Các yếu tố khác như vị trí địa lý, chất lượng xây dựng, tuổi đời của căn nhà, thiết kế nội thất, hay tình trạng pháp lý cũng có thể ảnh hưởng đáng kể đến mức giá cuối cùng.

Biểu đồ cũng cho thấy phần lớn dữ liệu tập trung ở vùng diện tích từ 1000 đến 2500 ft² và giá bán từ 100000 đến 300000 USD – đây có thể được xem là phân khúc nhà ở phổ biến nhất trong thị trường mẫu này.

4.7. Tổng kết khai phá và định hướng bài toán dự đoán.

Qua quá trình khai phá dữ liệu, có thể nhận thấy một số yếu tố có tương quan mạnh đến giá bán nhà như: diện tích sinh hoạt (GrLivArea), chất lượng tổng thể (OverallQual), diện tích gara (GarageArea), năm xây dựng (YearBuilt) và chất lượng bếp (KitchenQual). Những đặc trưng này đều thể hiện mối quan hệ trực tiếp hoặc gián tiếp với giá nhà, và có khả năng đóng vai trò quan trọng trong việc xây dựng mô hình dự đoán.

Từ đó, bài toán đặt ra là: **Dự đoán giá bán của một căn nhà tại Ames, Iowa dựa trên các đặc trưng mô tả vật lý, chất lượng và điều kiện sử dụng của căn nhà.**

Để giải quyết bài toán, nhóm sẽ tiến hành lựa chọn đặc trưng phù hợp và áp dụng các mô hình học máy như Linear Regression, Random Forest và Polynomial Regression để đánh giá hiệu quả dự đoán.

5. Thực nghiệm dự đoán giá nhà sử dụng học máy (ML).

Trong phần này, nhóm sẽ chia bộ dữ liệu làm 3 phần huấn luyện (train), tập kiểm tra (validation) và tập kiểm tra cuối cùng (testing) với tỉ lệ tương ứng 60% - 20% - 20% dựa trên phương pháp lấy mẫu ngẫu nhiên.

	Độ lớn tập dữ liệu
x_train, y_train	876
x_validation, y_validation	292
x_test, y_test	292

Bảng 12: Mô tả bộ dữ liệu sau khi được chia

Sau đó sử dụng 3 mô hình học máy để train và test tập dữ liệu:

1. Linear Regression.
2. Random Forest Regressor.
3. Polynomial Linear Regression.

Sau cùng, tính toán các chỉ số đánh giá hiệu suất mô hình dự báo, MAE, MSE, RMSE, $R^2 Score$.

5.1. Lựa chọn feature.

Khi dữ liệu có quá nhiều đặc trưng (feature), mô hình học máy dễ gặp phải hiện tượng overfitting. Do đó, việc lựa chọn ra những đặc trưng quan trọng và có ảnh hưởng mạnh nhất đến biến mục tiêu là một bước quan trọng nhằm giảm chiều dữ liệu, tăng hiệu quả và độ chính xác của mô hình.

Trong dự án này, sử dụng phương pháp SelecKBest, một kỹ thuật chọn lọc đặc trưng dựa trên thống kê, được hỗ trợ bởi thư viện Sklearn. SelectKBest sẽ tính điểm cho từng đặc trưng dựa vào mối quan hệ của nó với biến mục tiêu, sau đó giữ lại K đặc trưng có điểm cao nhất.

Bằng phương pháp SelectKBest, 30 feature được chọn là: ‘Neighborhood', 'OverallQual', 'YearBuilt', 'ExterQual', 'BsmtQual', TotalBsmtSF', 'GrLivArea', 'KitchenQual', 'GarageCars', 'GarageArea'.

5.2. Các mô hình học máy.

5.2.1. Mô hình Linear Regression.

Linear Regression là một trong những mô hình đơn giản nhưng hiệu quả trong bài toán hồi quy. Mô hình tìm ra mối quan hệ tuyến tính giữa biến đầu ra (giá nhà) và các biến đầu vào (đặc trưng), bằng cách ước lượng một hàm có dạng:

$$\widehat{y} = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + \mathrm{b}$$

Trong đó:

- $\widehat{y}$: giá nhà dự đoán.
- $x_1, x_2, \ldots, x_n$: các đặc trưng đầu vào.
- w_1, w_2, w_n : trọng số mà mô hình học được.
- b: sai số chênh lệch (bias).

Ưu điểm của Linear Regression là dễ triển khai, dễ diễn giải và chạy nhanh, đặc biệt phù hợp khi mối quan hệ giữa đầu vào và đầu ra gần như tuyến tính. Tuy nhiên, nhược điểm của mô hình là độ chính xác thường không cao khi dữ liệu có quan hệ phi tuyến hoặc chứa nhiều nhiễu. Do đó, Linear Regression thường được

dùng như một mô hình cơ sở (baseline) để so sánh với các mô hình phức tạp hơn như Random Forest hoặc XGBoost.

5.2.2. Mô hình Polynomial Linear Regression.

Polynomial Linear Regression (Hồi quy tuyến tính đa thức) là một phương pháp mở rộng của mô hình Linear Regression nhằm giải quyết các bài toán có mối quan hệ phi tuyến giữa đầu vào và đầu ra. Thay vì chỉ sử dụng các đặc trưng đầu vào ban đầu, mô hình này tạo thêm các đặc trưng mới bằng cách nâng lũy thừa các biến đầu vào (ví dụ: x^2,x^3 …), giúp mô hình có khả năng học các quan hệ phức tạp hơn.

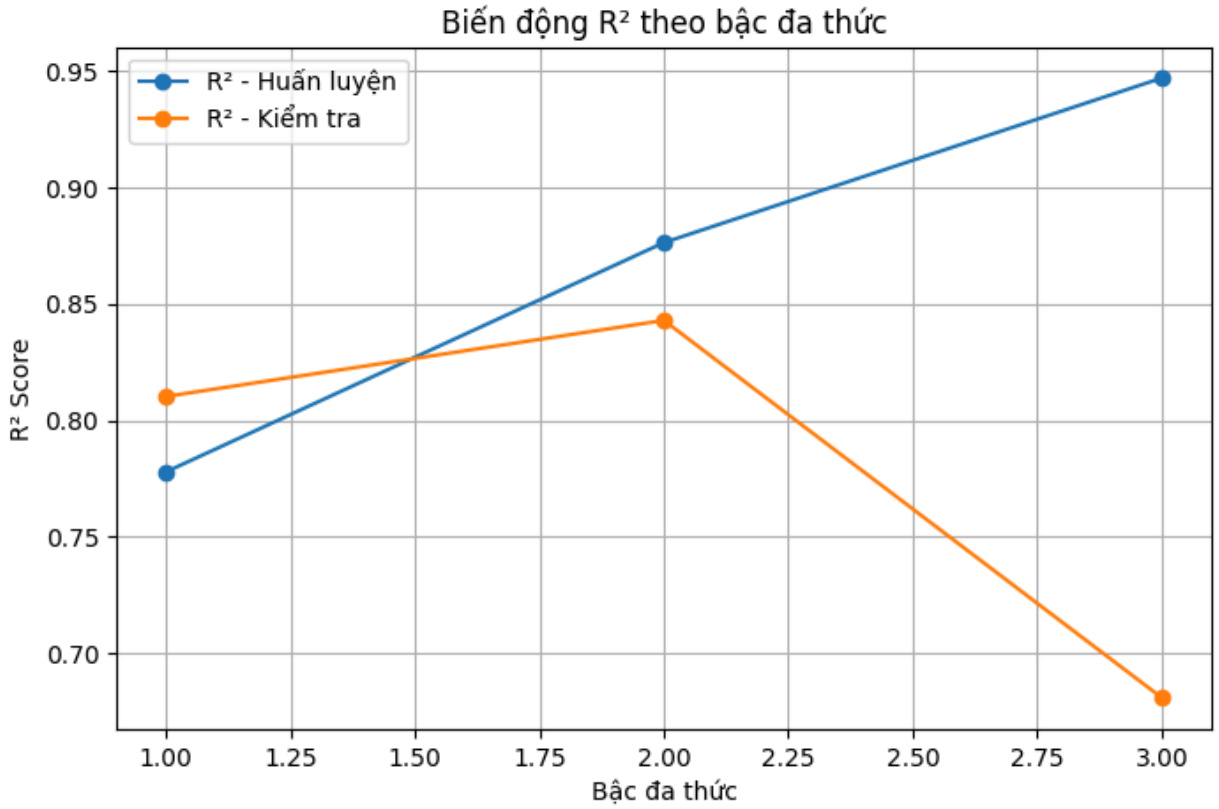
Phương trình của mô hình có dạng:

$$y = w_0 + w_1 x + w_1 x^2 + \ldots + w_n x^n + b$$

Trong đó:

- x là đặc trưng đầu vào, x^2, x^3, ..., x^n là các đặc trưng bậc cao.
- w_i là các trọng số mà mô hình học được.
- b là sai số chênh lệch (bias)

Trong quá trình thử nghiệm mô hình này với các bậc đa thức khác nhau. Khi tăng bậc của đa thức, mô hình trở nên phức tạp hơn và học kỹ dữ liệu huấn luyện, nên R^2 trên tập huấn luyện tăng. Tuy nhiên, do học cả nhiễu và đặc điểm không quan trọng, mô hình mất khả năng tổng quát hóa với dữ liệu mới, khiến R^2 trên tập kiểm tra giảm. Đây là biểu hiện rõ của hiện tượng overfitting trong học máy.



Hình 8: Biến động R² theo bậc đa thức trong mô hình Polynomial Linear Regression

Vì vậy ở mô hình Polynomial Linear Regression, chọn bậc của đa thức là 2.

5.2.3. Mô hình Random Forest.

Random Forest là một mô hình học máy thuộc nhóm thuật toán Ensemble Learning, kết hợp nhiều cây quyết định (Decision Trees) để cải thiện độ chính xác và độ ổn định của dự đoán. Thay vì chỉ dựa vào một cây duy nhất, Random Forest xây dựng hàng trăm hoặc hàng ngàn cây khác nhau bằng cách lấy mẫu ngẫu nhiên từ dữ liệu và đặc trưng đầu vào (bootstrapping và feature sampling). Kết quả dự đoán cuối cùng là trung bình (đối với bài toán hồi quy) hoặc bầu chọn đa số (đối với bài toán phân loại) từ các cây thành phần.

Trong bài toán dự đoán giá nhà, Random Forest có khả năng mô hình hóa các mối quan hệ phi tuyến phức tạp, tự động xử lý các tương tác giữa các đặc trưng mà không cần chuẩn hóa dữ liệu. Ngoài ra, mô hình còn có tính chống overfitting tốt nhờ vào sự ngẫu nhiên trong quá trình xây cây.

5.3. Các chỉ số đánh giá mô hình hồi quy.

Để đánh giá độ chính xác của các mô hình được sử dụng cho bài toán dự đoán giá nhà, các chỉ số được sử dụng bao gồm.

MAE (Mean Absolute Error): Trung bình tuyệt đối của sai số giữa giá trị dự đoán và giá trị thực tế. MAE đo lường mức sai lệch trung bình mà không quan tâm đến hướng sai số.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - \widehat{y_i} |$$

Trong đó: n là số lượng mẫu dữ liệu (số quan sát), y_i là giá trị thực tế, $\widehat{y_i}$ là giá trị dự đoán

MSE (Mean Squared Error): Trung bình bình phương của sai số. MSE phạt mạnh hơn đối với các sai số lớn, do đó nhạy với ngoại lệ.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

Trong đó: n là số lượng mẫu dữ liệu (số quan sát), y_i là giá trị thực tế, $\widehat{y_i}$ là giá trị dự đoán

RMSE (Root Mean Squared Error): Căn bậc hai của MSE, giúp đưa sai số về cùng đơn vị với biến đầu ra.

$$\text{RMSE} = \sqrt{MSE}$$

R^2 Score (Hệ số xác định): Đo lường mức độ mô hình giải thích được phương sai của dữ liệu. R^2 càng gần 1 thì mô hình càng tốt.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Trong đó: SS_{res} là tổng bình phương phần dư, SS_{tot} là tổng bình phương sai lệch so với trung bình.

5.4. Kết quả và đánh giá.

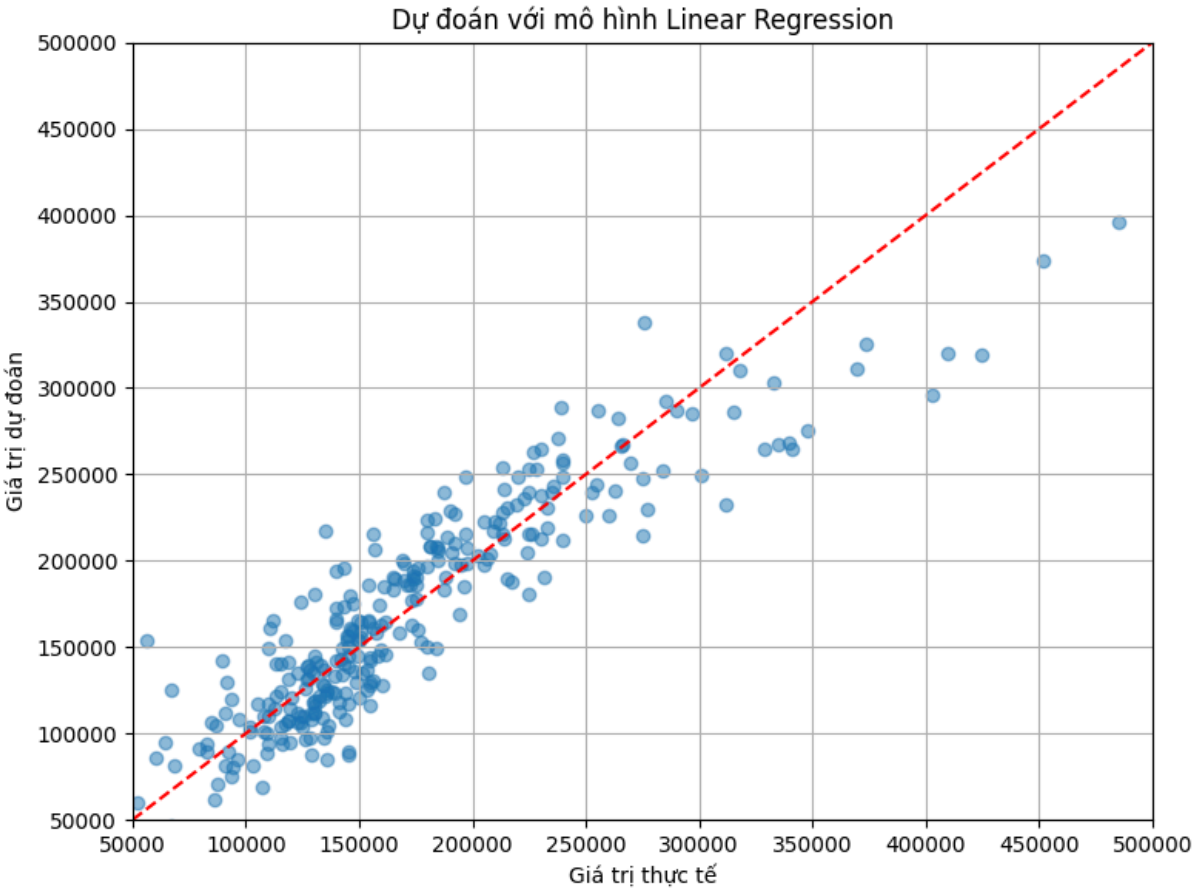
Từ kết quả tổng hợp từ đánh giá tập Test của 3 mô hình:

Model	MAE	MSE	RMSE	R2 Score
Linear Regression	22235.35	8.66×10^8	29435.25	0.8228
Random Forest	15425.31	4.65×10^8	21571.18	0.9048
Polynomial Linear Regression	20989.03	8.60×10^8	29330.68	0.8241

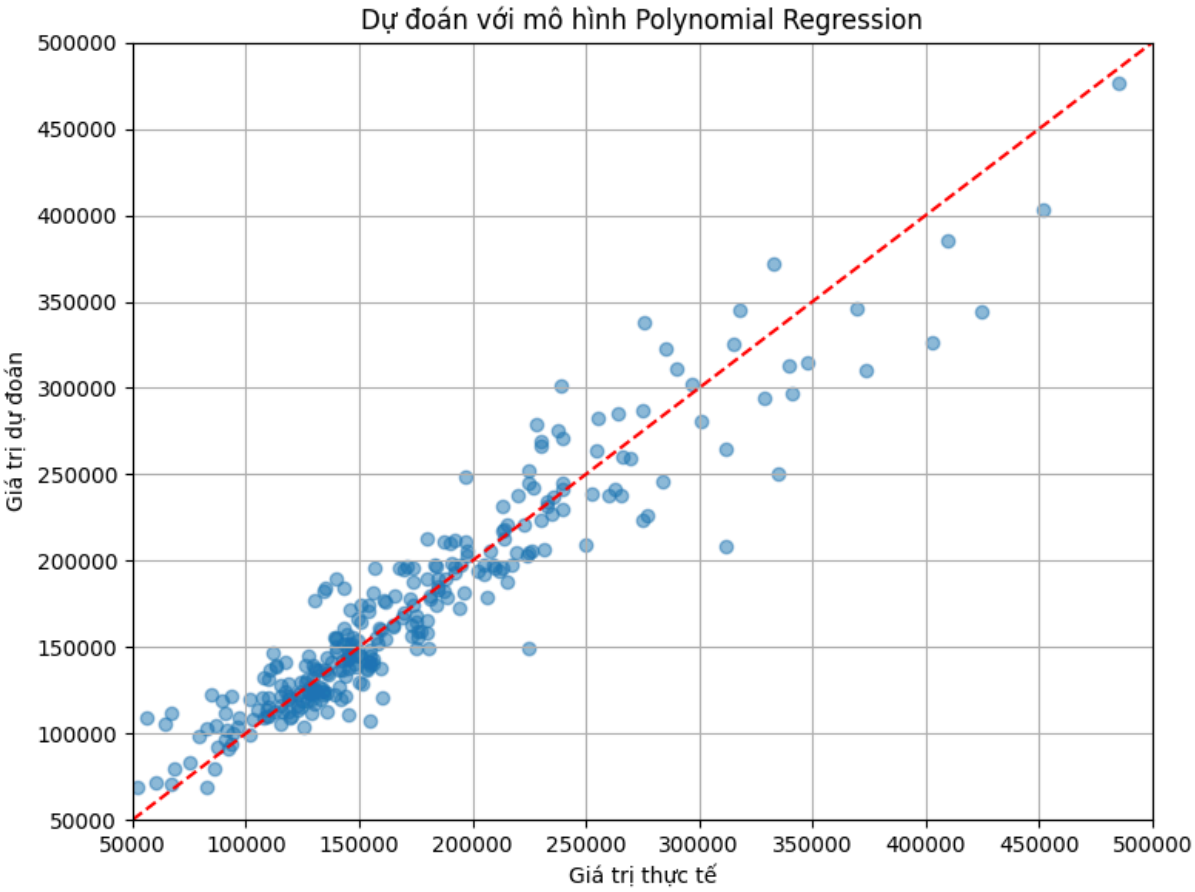
Bảng 13: Các chỉ số đánh giá cho từng mô hình

Kết quả đánh giá cho thấy mô hình Random Forest có hiệu suất vượt trội với MAE thấp nhất (15425.31), RMSE nhỏ nhất (21571.18) và hệ số R^2 cao nhất (0.9048).

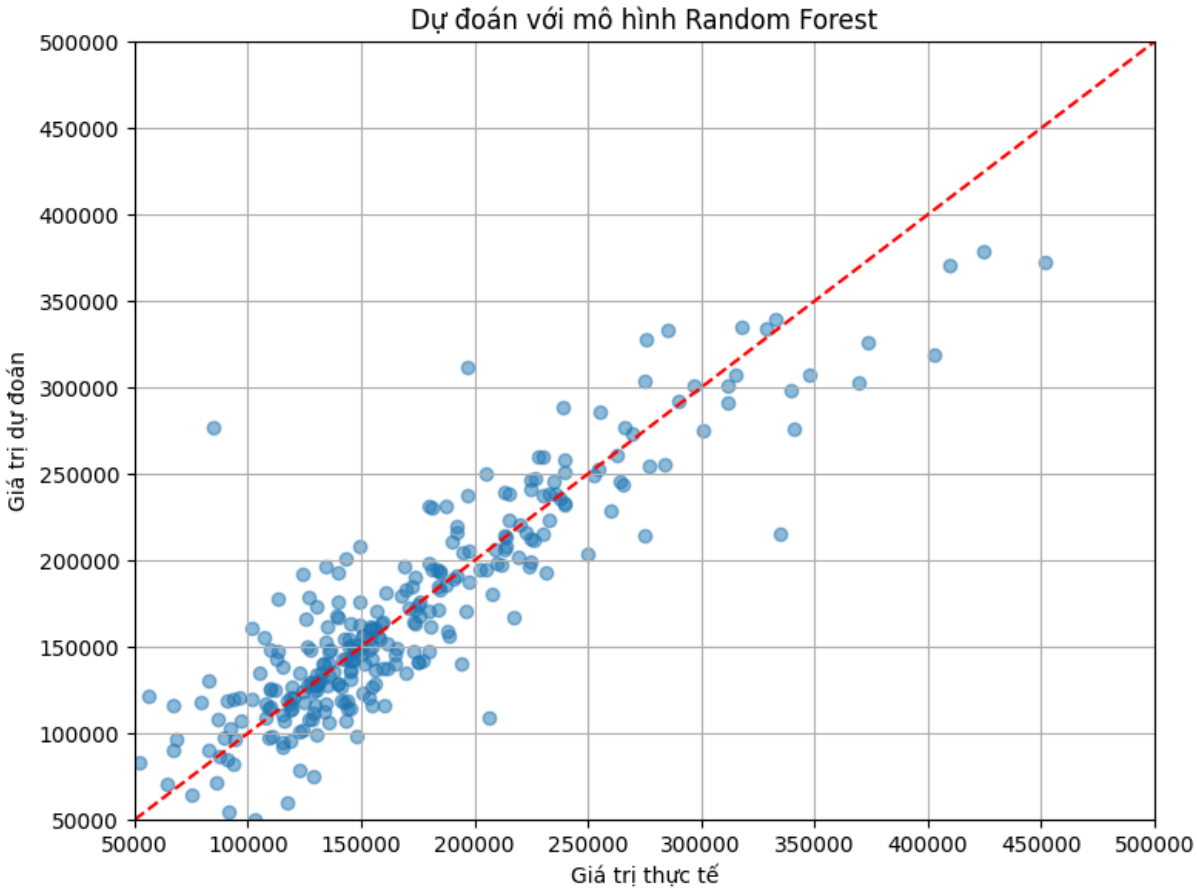
Để quan sát rõ hơn giá trị thực tế so với giá trị dự đoán của các mô hình, dưới đây là các biểu đồ Scatter trực quan các dự đoán của mô hình. Nếu mô hình dự đoán tốt, các điểm sẽ nằm gần đường chéo y = x (đường gạch đứt màu đỏ).



Hình 9: Kết quả dự đoán của mô hình Linear Regression so với thực tế.



Hình 10: Kết quả dự đoán của mô hình Polynomial Regression so với thực tế.



Hình 11: Kết quả dự đoán của mô hình Random Forest so với thực tế.

Biểu đồ Linear Regression cho thấy mô hình dự đoán kém chính xác ở vùng giá cao, nhiều điểm nằm dưới đường lý tưởng. Polynomial Regression cải thiện rõ rệt, các điểm phân bố sát đường chéo hơn, đặc biệt trong khoảng từ 100.000 đến 300.000 USD. Random Forest có phân tán rộng hơn nhưng dự đoán khá cân bằng trên toàn bộ miền giá, cho thấy khả năng tổng quát hóa tốt và ổn định hơn với dữ liệu đa dạng.

Điều này là nhờ vào khả năng của Random Forest trong việc mô hình hóa các mối quan hệ phi tuyến phức tạp và tự động xử lý tương tác giữa các đặc trưng, nhờ sử dụng nhiều cây quyết định kết hợp lại (ensemble learning). Linear Regression, dù đơn giản và dễ triển khai, chỉ nắm bắt được các mối quan hệ

tuyến tính giữa đặc trưng và giá nhà nên hiệu quả dự đoán còn hạn chế. Polynomial Linear Regression đã cải thiện phần nào nhờ mô hình hóa quan hệ phi tuyến bậc hai, nhưng sự cải thiện không đáng kể vì bản chất dữ liệu không đơn thuần là tuyến tính hay bậc hai, mà có tính phi tuyến và tương tác cao. Do đó, Random Forest là lựa chọn phù hợp và hiệu quả nhất trong ba mô hình được thử nghiệm

KẾT LUẬN

Dự án "Phân tích các yếu tố ảnh hưởng đến giá nhà ở Ames, Iowa (Hoa Kỳ)" đã cung cấp một cái nhìn toàn diện và có hệ thống về cách các yếu tố đặc trưng của bất động sản ảnh hưởng đến giá bán. Bằng cách áp dụng quy trình phân tích dữ liệu hiện đại từ thu thập, tiền xử lý, khám phá dữ liệu đến xây dựng và đánh giá mô hình học máy, nhóm đã rút ra được nhiều kết luận có giá trị thực tiễn.

Trên cơ sở khai thác tập dữ liệu phong phú từ Kaggle, nhóm xác định được những yếu tố có ảnh hưởng mạnh nhất đến giá nhà bao gồm: diện tích sinh hoạt (GrLivArea), chất lượng tổng thể của ngôi nhà (OverallQual), diện tích gara (GarageArea), năm xây dựng (YearBuilt), và chất lượng bếp (KitchenQual). Đây đều là những đặc trưng phản ánh rõ ràng mức độ tiện nghi, không gian sử dụng và giá trị đầu tư của bất động sản.

Ba mô hình học máy đã được triển khai để thực hiện dự đoán: Linear Regression, Polynomial Linear Regression và Random Forest. Trong đó:

- Linear Regression là mô hình đơn giản, dễ cài đặt nhưng có độ chính xác trung bình ($R^2 \approx 0.82$), thích hợp để làm mô hình cơ sở (baseline).
- Polynomial Linear Regression giúp mô hình hóa tốt hơn các quan hệ phi tuyến, tuy nhiên dễ bị quá khớp dữ liệu (overfitting).
- Random Forest cho kết quả tốt nhất với độ chính xác cao ($R^2 \approx 0.90$), sai số thấp, và khả năng khái quát tốt nhờ cơ chế tổng hợp nhiều cây quyết định.

Tuy nhiên, hạn chế vẫn tồn tại như sự thiếu hụt dữ liệu trong một số cột, hoặc ảnh hưởng của các giá trị ngoại lệ khiến mô hình có thể bị lệch. Do đó, trong các nghiên cứu mở rộng, cần thử nghiệm thêm các mô hình tiên tiến hơn như

XGBoost hoặc LightGBM, kết hợp tinh chỉnh siêu tham số và chiến lược chọn lọc đặc trưng tối ưu.

Tổng thể, dự án không chỉ giúp nhóm hiểu sâu hơn về thị trường bất động sản qua dữ liệu, mà còn nâng cao kỹ năng thực hành phân tích dữ liệu, trực quan hóa và triển khai mô hình học máy – những kỹ năng quan trọng trong kỷ nguyên dữ liệu hiện nay.

TÀI LIỆU THAM KHẢO

1. Kaggle. House Prices - Advanced Regression Techniques Dataset. Truy cập tại: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

2. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

3. Waskom, M. (2020). Seaborn: Statistical Data Visualization. Truy cập tại: https://seaborn.pydata.org/

4. McKinney, W. (2010). Data Analysis with Python and Pandas. Truy cập tại: https://pandas.pydata.org/

5. Các tài liệu và bài giảng môn "Lập trình phân tích dữ liệu với Python" – Giảng viên ThS. Nguyễn Văn Thiệu – Khoa CNTT, Trường Đại học Phenikaa.