Dự đoán giá nhà sử dụng mô hình máy học

Nguyễn Đăng Khoa, Lê Thị Trúc Ly, Lê Đoàn Kim Ngân, Lâm Tú Nhi

Giới thiệu

Định nghĩa vấn đề

- **Input:** Thông tin tổng quát, đặc điểm vât lí, chất lượng và tình trạng của ngôi nhà
- **Output:** Giá nhà dự đoán gần với giá thực tế nhất

Thách thức

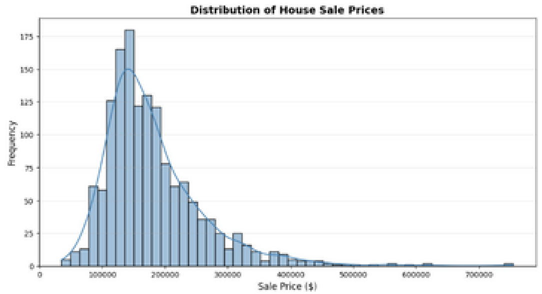
- Biến mục tiêu (SalePrice) lệch phải mạnh
- Dữ liệu thiếu phức tạp: nhiều giá trị NaN mang nghĩa "không có"

Mục tiêu : Xây dựng hệ thống dự đoán giá nhà chính xác, ổn định và có khả năng tổng quát tốt.

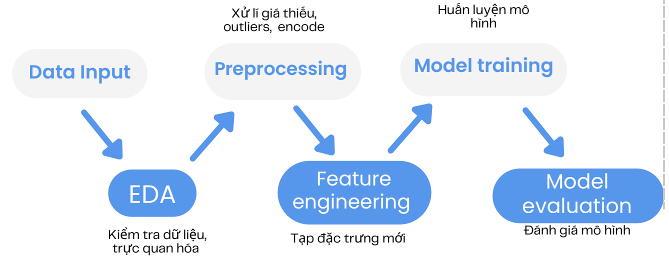
Tập dữ liệu

- Nguồn dữ liệu: House Prices - Advanced Regression Techniques
- Số lượng: 1460 mẫu (training set), 1459 mẫu (test set) và 81 đặc trưng

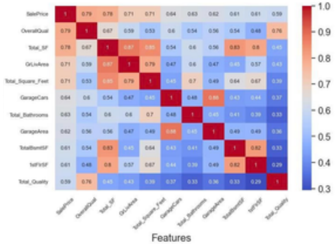
https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview



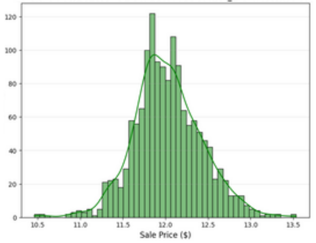
Phương pháp đề xuất



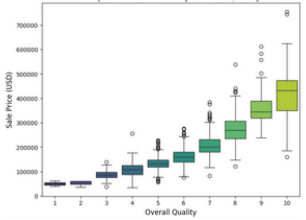
1.Phân tích, khám phá dữ liệu (EDA)



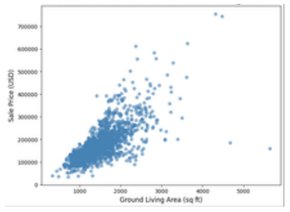
Top 10 đặc trưng có tương quan cao nhất



SalePrice sau khi log transform

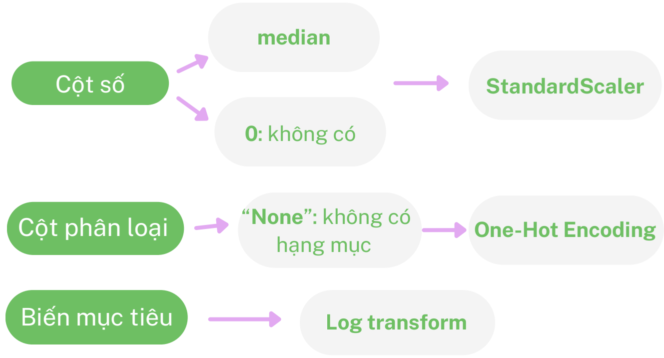


Chất lượng càng cao thì giá bán càng tốt



Giá nhà phụ thuộc nhiều vào diện tích đất ở

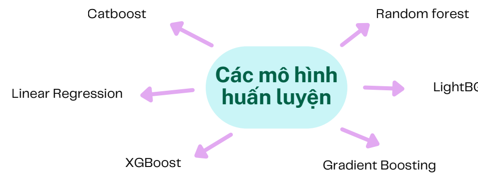
2.Tiền xử lí dữ liệu



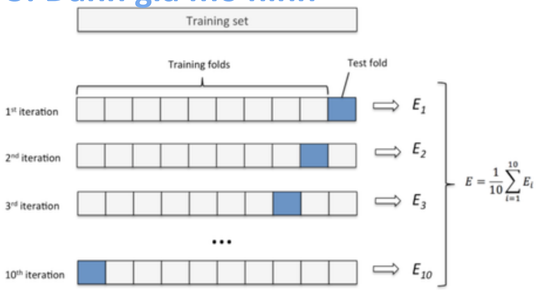
3. Kỹ thuật tạo đặc trưng

TotalSF	Tổng diện tích sàn (tầng trệt + tầng 2 + tầng hầm)	HasPool	Có hồ bơi
TotalPorchSF	Tổng diện tích hiên / ban công	HasFireplace	Có lò sưởi
TotalBath	Tổng số phòng tắm (tính cả tầng hầm, quy đổi	HasGarage	Có gara
TotalRooms	Tổng số phòng sinh hoạt (bao gồm phòng tắm)	HasBsmt	Có tầng hầm
TotalSpace	Tổng không gian sử dụng (diện tích nhà + garage +	HasPorch	Có hiên/ ban côn

4.Huấn luyện mô hình



5. Đánh giá mô hình



KFold chia dữ liệu thành 5 phần bằng nhau , trong đó 4 tập train và 1 tập test

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

Đo độ lệch trung bình giữa giá trị dự đoán và giá trị thực tế

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \longrightarrow$$

Đo mức độ mà mô hình có thể giải thích được sự biến thiên của dữ liệu thực tế

Thực nghiệm

- **Lần 1:** Baseline: xử lí giá trị thiếu, log trasform `SalePrice`
- **Lần 2:** Tạo thêm các đặc trưng mới (TotalSF, TotalPorchSF,..)
- **Lần 3:** Kết hợp các mô hình

Lần 1: Baseline: xử lí giá trị thiếu, log trasform `SalePrice`

Model	RSME	Kaggle score
XGBoost	**0.1287 ± 0.0176**	**0,12364**
CatBoost	0.1255 ± 0.0174	0,12416
Gradient Boosting	0.1324 ± 0.0211	0,12862
LightGBM	0.1335 ± 0.0192	0,12896

Lần 2: Tạo thêm các đặc trưng mới (TotalSF, TotalPorchSF,..)

Model	RSME	Kaggle score
XGBoost	0.1299 ± 0.0155	0,1232
CatBoost	**0.1251 ± 0.0162**	**0,12188**
Gradient Boosting	0.1299 ± 0.0155	0,1232
LightGBM	0.1343 ± 0.0208	0,12885

Lần 3: Kết hợp các mô hình

Model	RSME	Kaggle score
Stacking CB+RF+GB	0.1255 ± 0.0186	0,12247
Weight CB+XGB+LGB	**0.1256 ± 0.0179**	**0,12091**
Weight LGB+XGB+GB	0.1279 ± 0.0190	0,12318

Kết luận

Mô hình đạt kết quả trên Kaggle tốt
→ Weight Avarage CatBoost + XGBoost + LightGBM
-> 0.12091