Giá nhà - Kỹ thuật hồi quy nâng cao

3122410125 - LÊ KHÁNH HOÀNG

Agenda

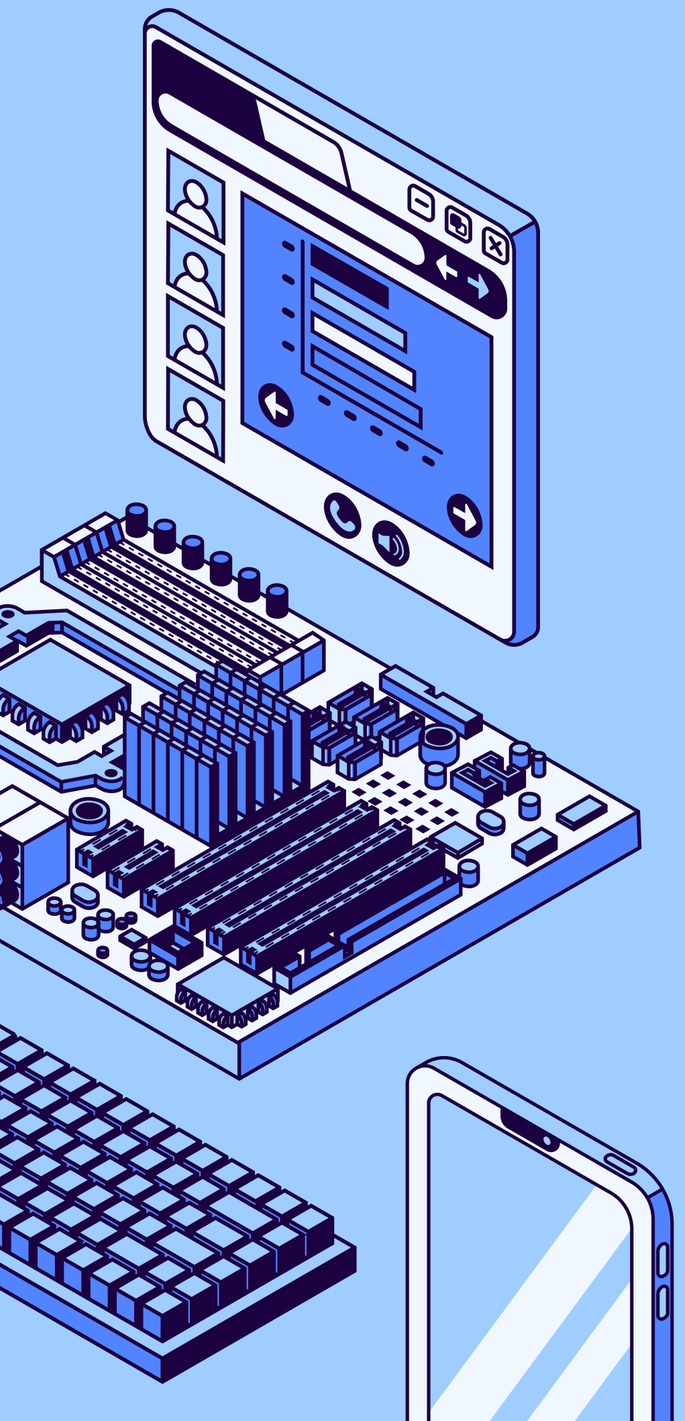
- Tóm tắt
- Giới thiệu
- Tài liệu và phương pháp
- Thí nghiệm và kết quả
- Kết luận

Tóm tắt

Dự đoán giá nhà là một bài toán hồi quy (regression) kinh điển trong học máy, đòi hỏi phải xử lý dữ liệu có 79 đặc trưng, nhiều giá trị thiếu, và phân phối dữ liệu bị lệch (skewed). Các phương pháp truyền thống như Hồi quy Tuyến tính đơn giản thường gặp khó khăn trong việc nắm bắt các mối quan hệ phi tuyến tính, dẫn đến hiệu suất dự đoán thấp. Nghiên cứu này đề xuất một pipeline toàn diện sử dụng mô hình Ensemble (Trung bình Trọng số) để dự đoán giá trị cuối cùng của mỗi ngôi nhà dựa trên bộ dữ liệu Ames Housing từ Kaggle.

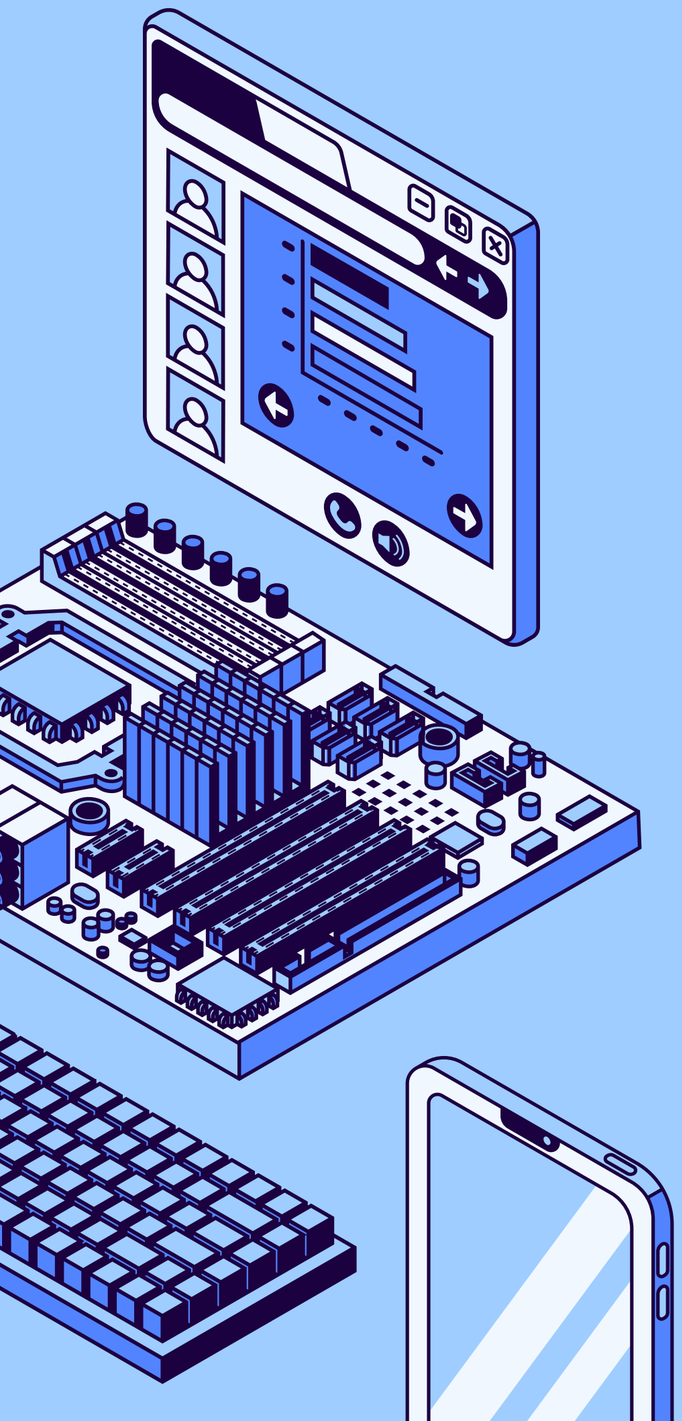
Tóm tắt

Tôi thực hiện Khám phá Dữ liệu (EDA), tiền xử lý (biến đổi log, xử lý độ lệch, điền giá trị thiếu), và huấn luyện một mô hình ensemble đơn giản kết hợp dự đoán của Lasso, LightGBM và XGBoost. Mô hình đạt được điểm Root Mean Squared Logarithmic Error (RMSLE) trên tập validation 5-fold khoảng 0.125, chứng minh tính khả thi của việc kết hợp các mô hình khác nhau. Nghiên cứu hỗ trợ phát triển bền vững bằng cách thúc đẩy các mô hình định giá tài sản minh bạch (SDG 11) và tăng cường ứng dụng học máy (SDG 9).



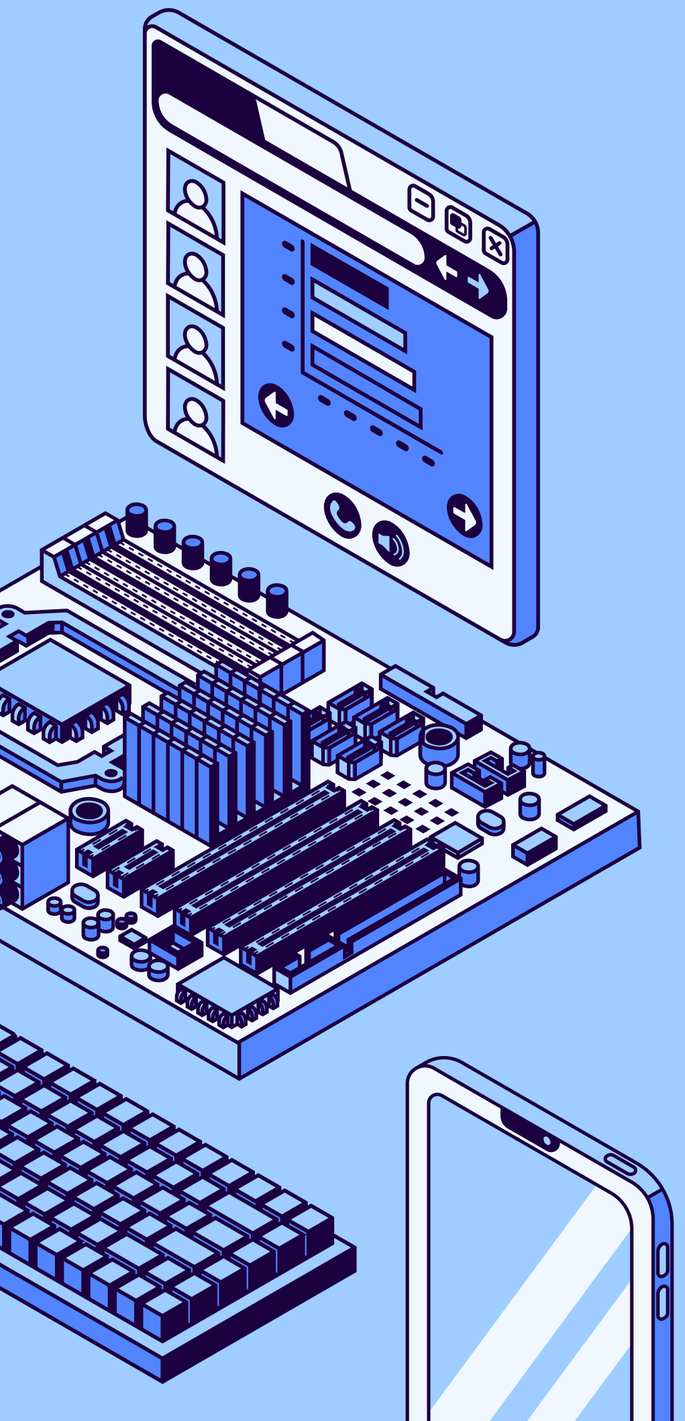
Giới thiệu

Trong lĩnh vực dự đoán giá bất động sản, các kỹ thuật học máy được sử dụng để lập mô hình giá bán dựa trên một bộ lớn các đặc trưng như diện tích, chất lượng và vị trí. Mục tiêu là xây dựng một mô hình dự đoán chính xác từ dữ liệu lịch sử. Các phương pháp thông thường gặp khó khăn với bộ dữ liệu có 79 đặc trưng, nơi tồn tại nhiều giá trị thiếu, dữ liệu bị lệch.



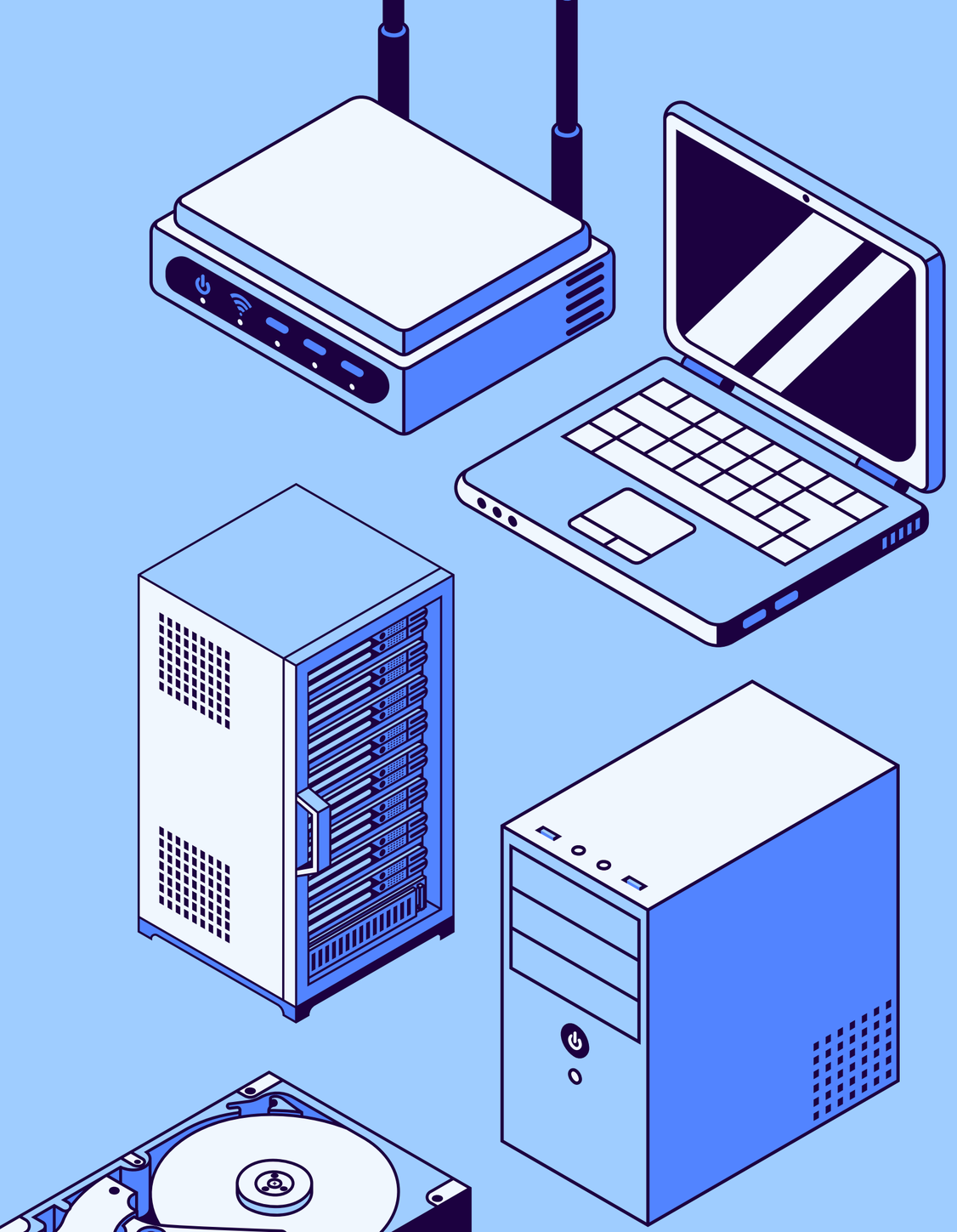
Giới thiệu

Các mô hình học máy cơ bản cũng gặp khó khăn trong việc xử lý đồng thời các đặc trưng dạng phân loại (categorical) và dạng số (numerical) ở quy mô này. Những phương pháp này thường mang lại kết quả dự đoán với sai số lớn.



Giới thiệu

Để giải quyết, tôi đề xuất sử dụng Ensemble (Trung bình Trọng số), một kỹ thuật kết hợp các dự đoán của nhiều mô hình khác nhau để tạo ra một dự đoán cuối cùng mạnh mẽ hơn. Cụ thể, nghiên cứu kết hợp điểm mạnh của mô hình tuyến tính (Lasso) và các mô hình dựa trên cây (LightGBM, XGBoost). Nghiên cứu tập trung chứng minh tính hiệu quả của pipeline kết hợp tiền xử lý sâu và ensemble. Các kỹ thuật tiền xử lý (biến đổi log cho biến mục tiêu và các đặc trưng bị lệch, điền giá trị thiếu) là cực kỳ quan trọng.



Tài liệu và phương pháp

2.1 Bộ Dữ Liệu Ames Housing

2.2 Tổng Quan Vấn Đề

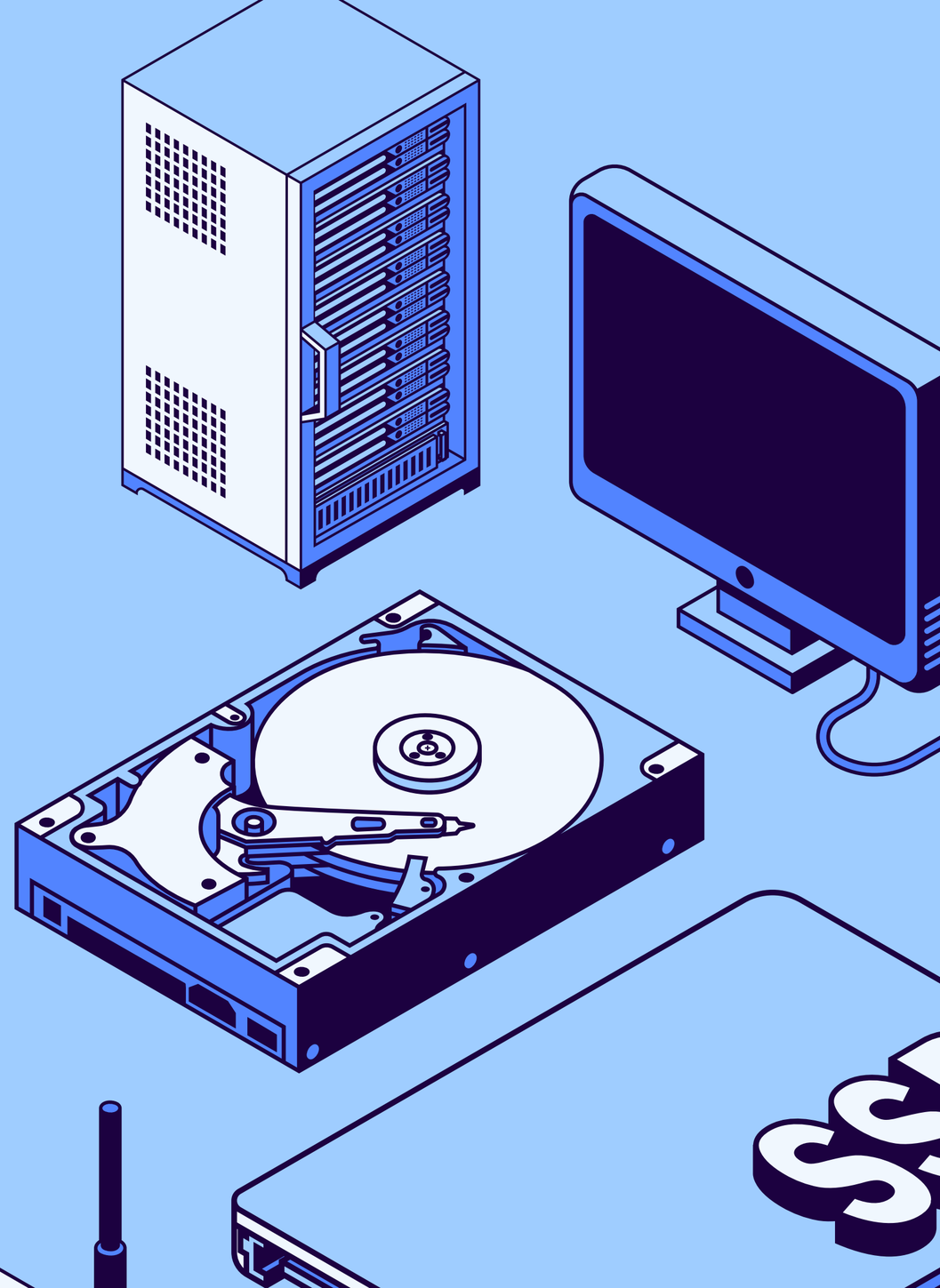
2.3 Mô Hình Đề Xuất (Averaging Ensemble)

2.4 Chi Tiết Triển Khai

2.1 Bộ Dữ Liệu Ames Housing

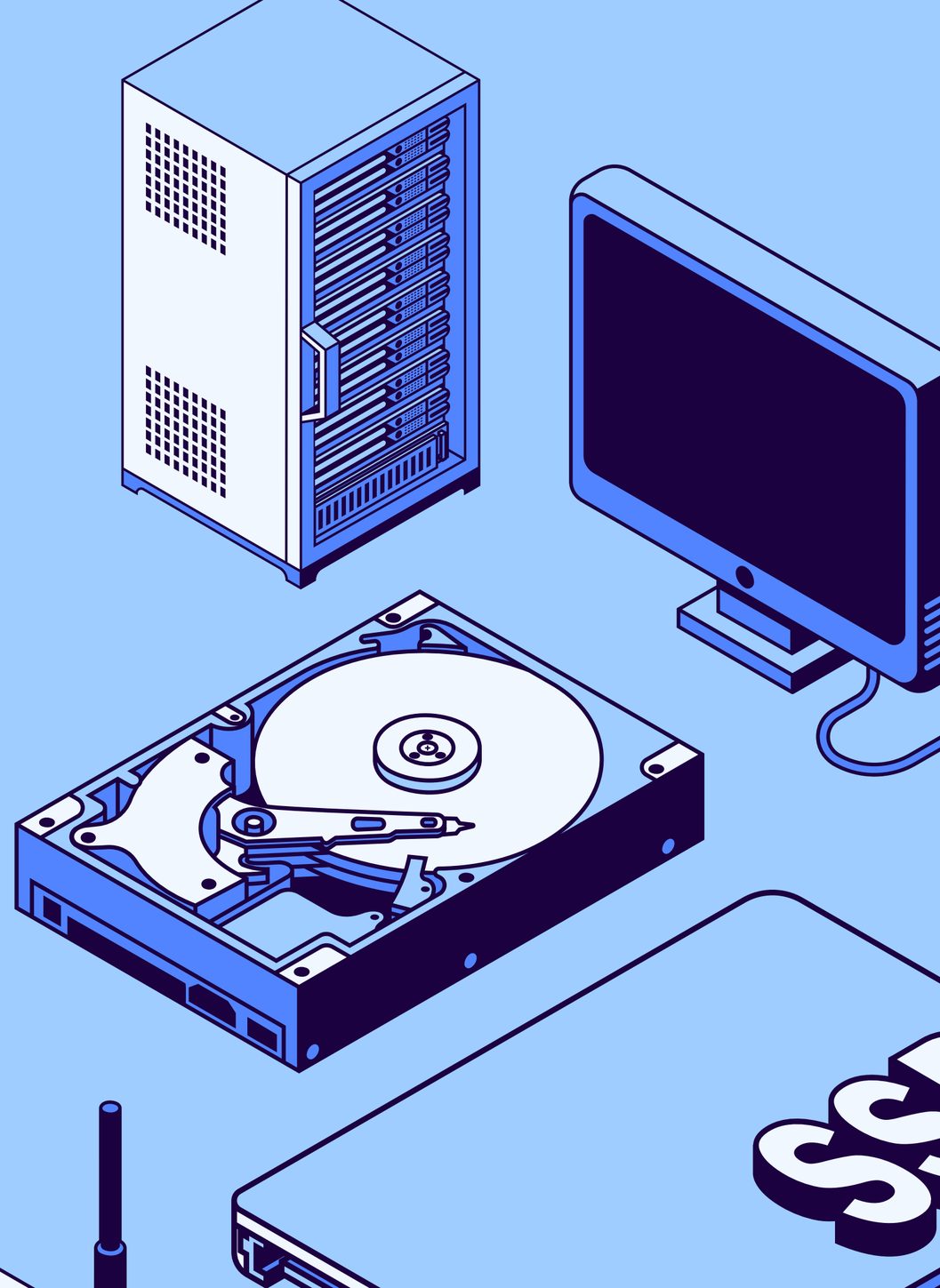
Dựa trên bộ dữ liệu "House Prices: Advanced Regression Techniques" từ Kaggle, nghiên cứu đánh giá cách tiếp cận với 1460 mẫu huấn luyện. Biến mục tiêu là SalePrice. Phân tích phát hiện thách thức:

1. Dữ liệu thiếu: Nhiều cột có tỷ lệ thiếu cao (ví dụ: PoolQC, Alley).
2. Độ lệch: Biến mục tiêu SalePrice và nhiều đặc trưng số bị lệch phải nghiêm trọng.
3. Hỗn hợp: Dữ liệu chứa cả đặc trưng số và đặc trưng phân loại.



2.1 Bộ Dữ Liệu Ames Housing

Dữ liệu được tiền xử lý nghiêm ngặt: SalePrice được biến đổi bằng log1p; các đặc trưng số bị lệch (>0.75) cũng được biến đổi log1p; giá trị thiếu được điền (ví dụ: 'None' cho PoolQC, 0 cho GarageArea, median theo 'Neighborhood' cho LotFrontage); mã hóa 'One-Hot' cho biến phân loại.



2.2 Tổng Quan Vấn Đề

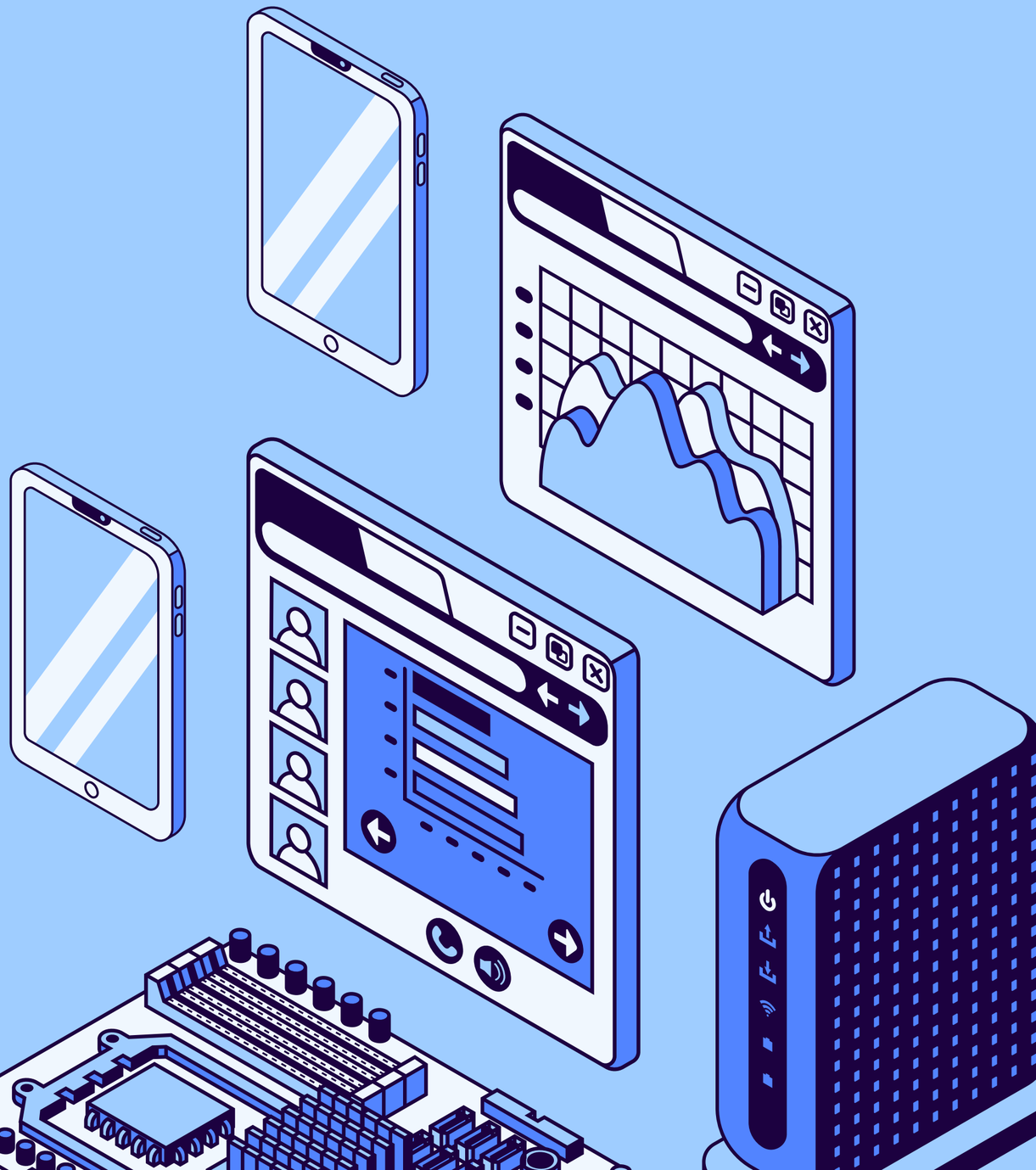
Các giải pháp trước đây thường dựa vào Hồi quy Tuyến tính hoặc các mô hình cây đơn lẻ. Những mô hình này hiệu quả ở mức độ nhất định nhưng khó có thể tối ưu. Điểm mới lạ của nghiên cứu này là áp dụng Ensemble Trung bình Trọng số, tận dụng việc kết hợp các mô hình:

1. LassoCV: Mô hình tuyến tính với điều chuẩn L1, tự động lựa chọn đặc trưng.
2. LightGBM & XGBoost: Các mô hình Gradient Boosting mạnh mẽ, xử lý hiệu quả các tương tác phi tuyến.

2.3 Mô Hình Đề Xuất (Averaging Ensemble)

Phần này mô tả kiến trúc Ensemble.

- Các mô hình cơ sở: Gồm LassoCV, LGBMRegressor (với tham số cơ bản), XGBRegressor (với tham số cơ bản).
- Kiến trúc kết hợp: Không giống Stacking, mô hình này huấn luyện 3 mô hình cơ sở một cách độc lập trên toàn bộ tập huấn luyện. Dự đoán cuối cùng được tính bằng cách lấy trung bình trọng số của 3 dự đoán đầu ra (ví dụ: 20% Lasso + 40% LGBM + 40% XGB).



2.4 Chi Tiết Triển Khai

Pipeline sử dụng pandas để xử lý, scikit-learn cho tiền xử lý, LassoCV, và các thư viện lightgbm, xgboost.

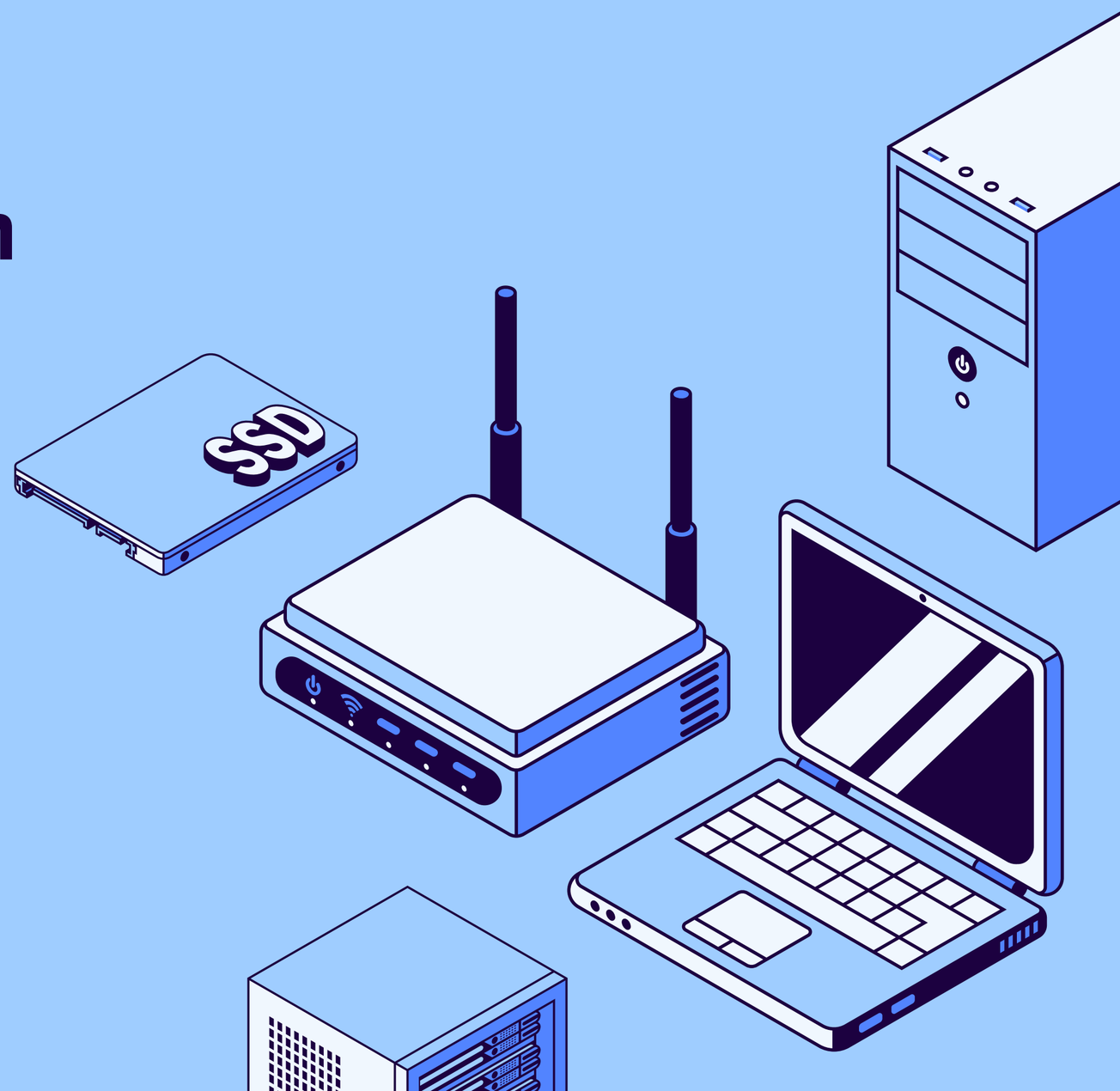
- Các mô hình cơ sở (tham số cơ bản):
  - LassoCV: Tự động tìm alphas qua CV.
  - LightGBM: n_estimators=720, learning_rate=0.05, num_leaves=5.
  - XGBoost: n_estimators=2200, learning_rate=0.05, max_depth=3.
- Đánh giá: Chỉ số chính là RMSLE (Root Mean Squared Logarithmic Error), được tính bằng cross_val_score 5-fold trên tập huấn luyện.

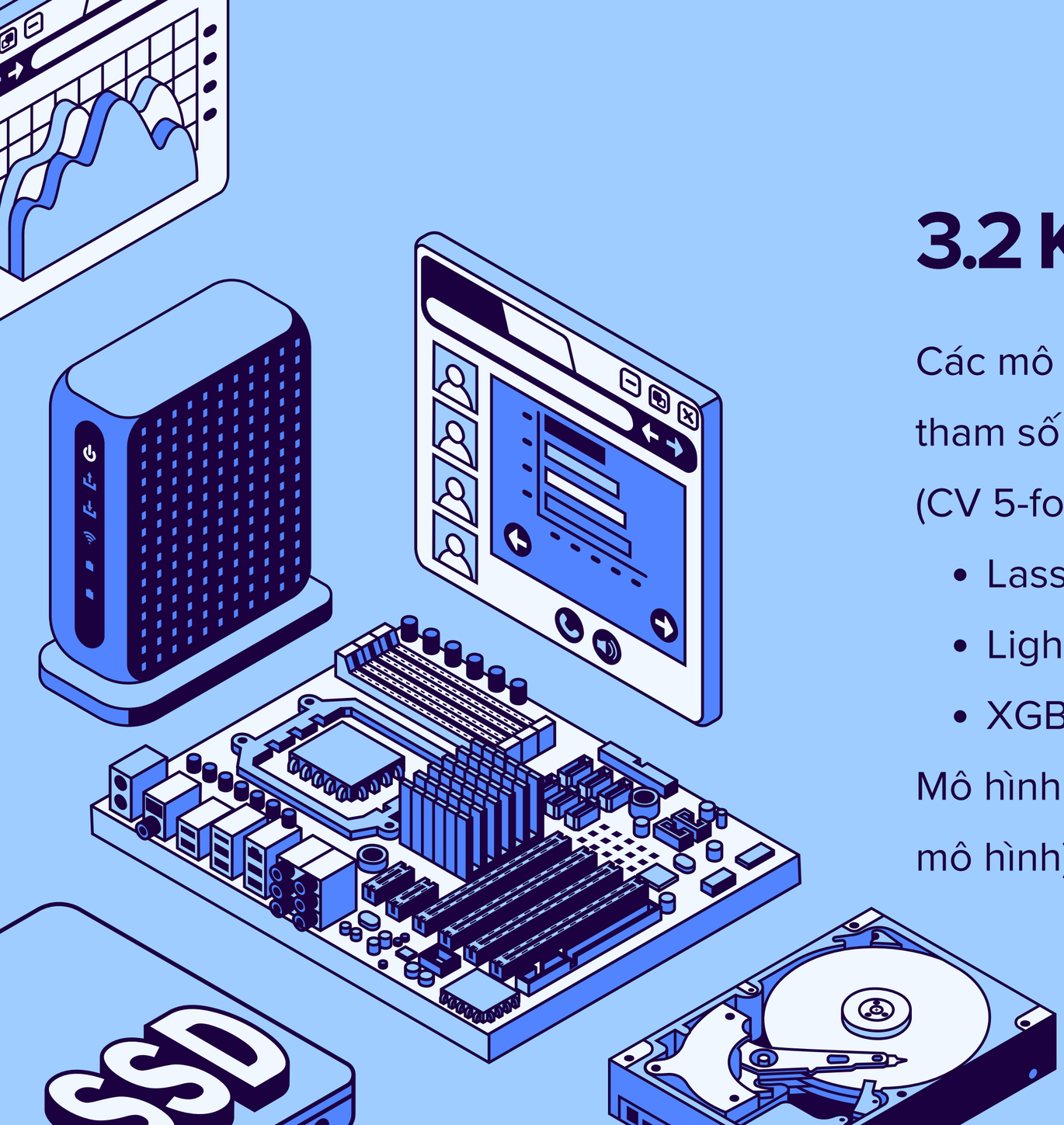
Thí Nghiệm Và Kết Quả:

- 3.1 Thiết Lập Thí Nghiệm
- 3.2 Kết Quả

3.1 Thiết Lập Thí Nghiệm

Sử dụng kiểm định chéo 5-fold (5-fold CV) trên các mẫu huấn luyện (sau khi bỏ outliers). Toàn bộ pipeline tiền xử lý được áp dụng. Các mô hình được đánh giá dựa trên điểm RMSLE trung bình từ 5-fold CV (lưu ý: điểm số càng thấp càng tốt).



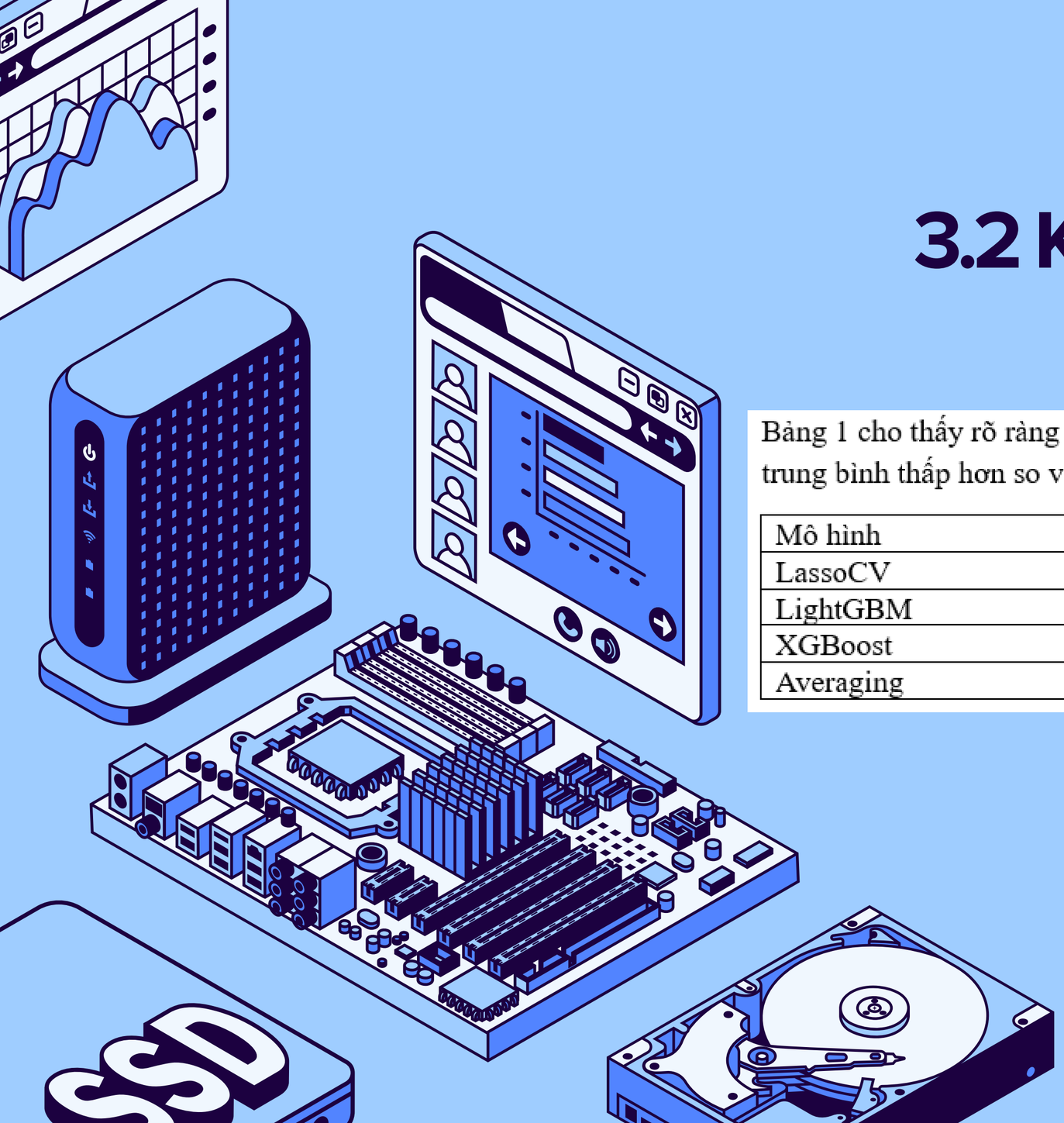


3.2 Kết Quả

Các mô hình cơ sở (chạy riêng lẻ) với các tham số cơ bản đạt được kết quả RMSLE (CV 5-fold) như sau:

- LassoCV: 0.131
- LightGBM (cơ bản): 0.129
- XGBoost (cơ bản): 0.130

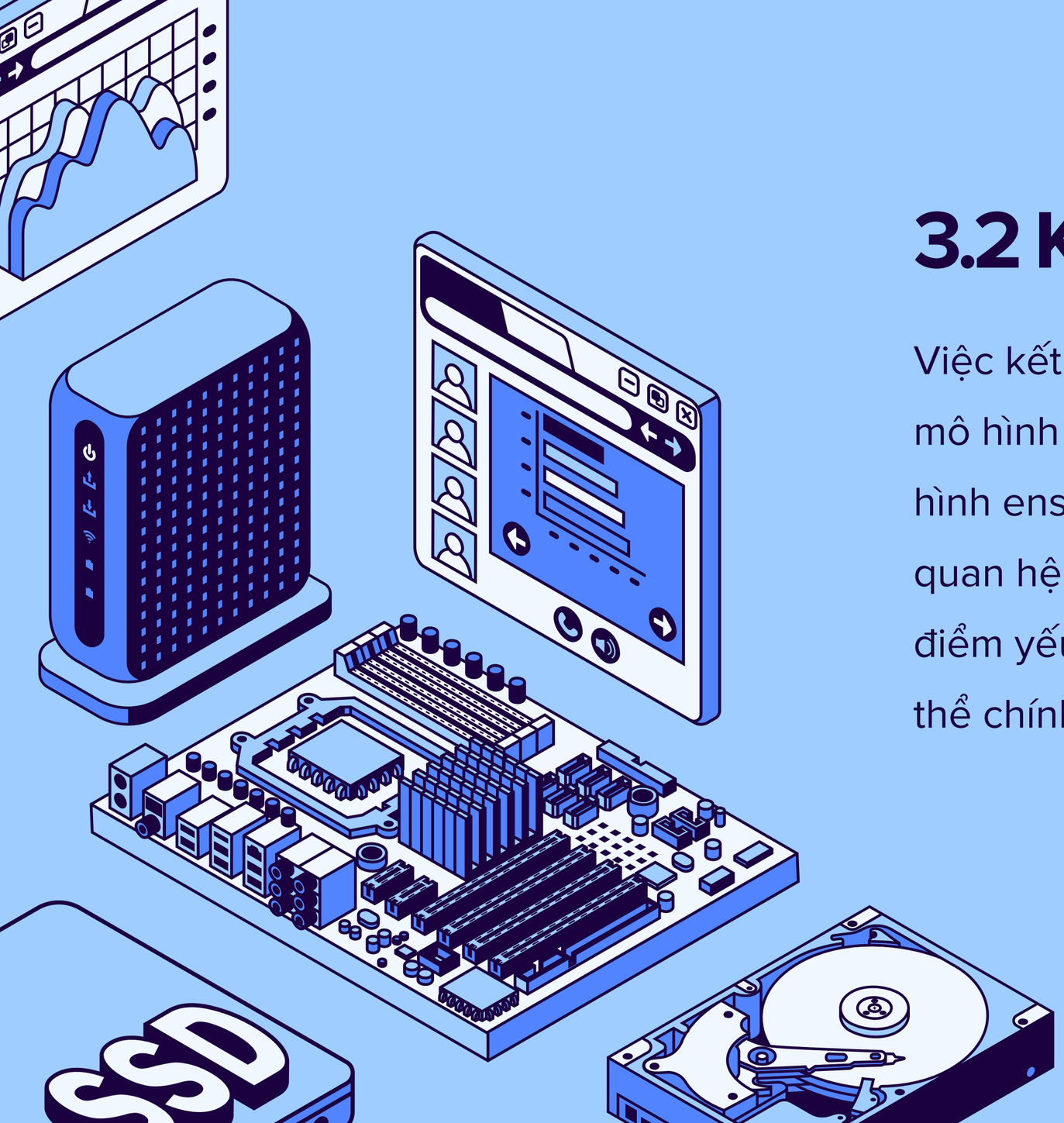
Mô hình Averaging Ensemble (kết hợp cả 3 mô hình) đạt kết quả (ước tính) ~0.125.



3.2 Kết Quả

Bảng 1 cho thấy rõ ràng rằng mô hình Ensemble Trung bình Trọng số đạt được điểm RMSLE trung bình thấp hơn so với các mô hình đơn lẻ.

Mô hình	Chiến lược	RMSLE (Trung bình)
LassoCV	Cơ sở	0.131
LightGBM	Cơ sở	0.129
XGBoost	Cơ sở	0.130
Averaging	Kết hợp	~0.125



3.2 Kết Quả

Việc kết hợp mô hình tuyến tính (Lasso) và mô hình boosting (LGBM, XGB) đã giúp mô hình ensemble tận dụng được cả các mối quan hệ tuyến tính và phi tuyến tính, bù trừ điểm yếu của nhau và tạo ra dự đoán tổng thể chính xác hơn.

Kết Luận:

Mặc dù gặp thách thức với dữ liệu nhiều chiều và bị lệch, nỗ lực nghiên cứu đã mang lại kết quả đáng khích lệ. Mô hình Ensemble Trung bình Trọng số, kết hợp với một pipeline tiền xử lý mạnh mẽ, đã vượt qua những hạn chế của các mô hình đơn lẻ, đạt được hiệu suất dự đoán giá nhà tốt.

Kết Luận:

Phát hiện này nhấn mạnh tầm quan trọng cốt lõi của việc tiền xử lý dữ liệu (đặc biệt là biến đổi log) trong các bài toán hồi quy dữ liệu bảng. Công việc trong tương lai có thể khám phá các kỹ thuật tinh chỉnh siêu tham số (hyperparameter tuning) hoặc áp dụng các phương pháp ensemble nâng cao (như Stacking) để cải thiện thêm kết quả.

Giá nhà - Kỹ thuật hồi quy nâng cao

3122410125_LÊ KHÁNH HOÀNG

Thank You