Dự đoán giá nhà

Seminar

Môn: Phân tích dữ liệu thông minh

Vũ Đăng Hoàng Long (18120203)
Nguyễn Huỳnh Đại Lợi (18120198)
Trần Thanh Phúc (18120225)

Huỳnh Long Nam (18120212)
Nguyễn Đăng Trung Tiến (18120591)

House Prices - Advanced Regression Techniques | Kaggle

Nội dung

1. Dữ liệu và bài toán
2. Khám phá dữ liệu
3. Thử nghiệm mô hình
4. Đánh giá
5. Tổng kết

Dữ liệu và bài toán

Dữ liệu

- Giá các căn hộ được bán tại Ames, Iowa từ 2006 đến 2010
- Gốc: 2930 dòng x 80 cột (Kaggle chỉ dùng 1460 dòng để train)
- 43 thuộc tính phân loại, 37 thuộc tính số

Bài toán

- Dự đoán giá nhà từ các thông tin cơ bản của một căn hộ (vị trí xây, diện tích, tiện nghi,...)
- Yêu cầu:
  - Kỹ năng xử lý dữ liệu
  - Kỹ năng sử dụng mô hình regression nâng cao

Khám phá dữ liệu

Đọc mô tả dữ liệu

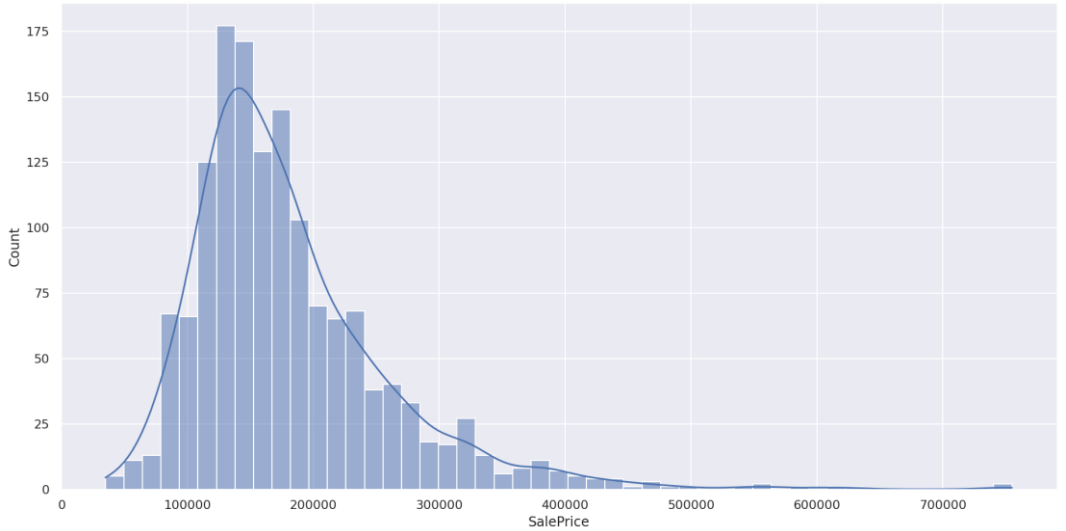
- Một số cột thiếu dữ liệu là có chủ đích:
  - Alley, BsmtQual, GarageType, PoolQC
- Một số cột có dạng multiple choice
  - Condition1/Condition2, Exterior1st/Exterior2nd,...

Thuộc tính phụ thuộc

- Dạng hình chuông
- Lệch phải → Có nên biến đổi logarithm?

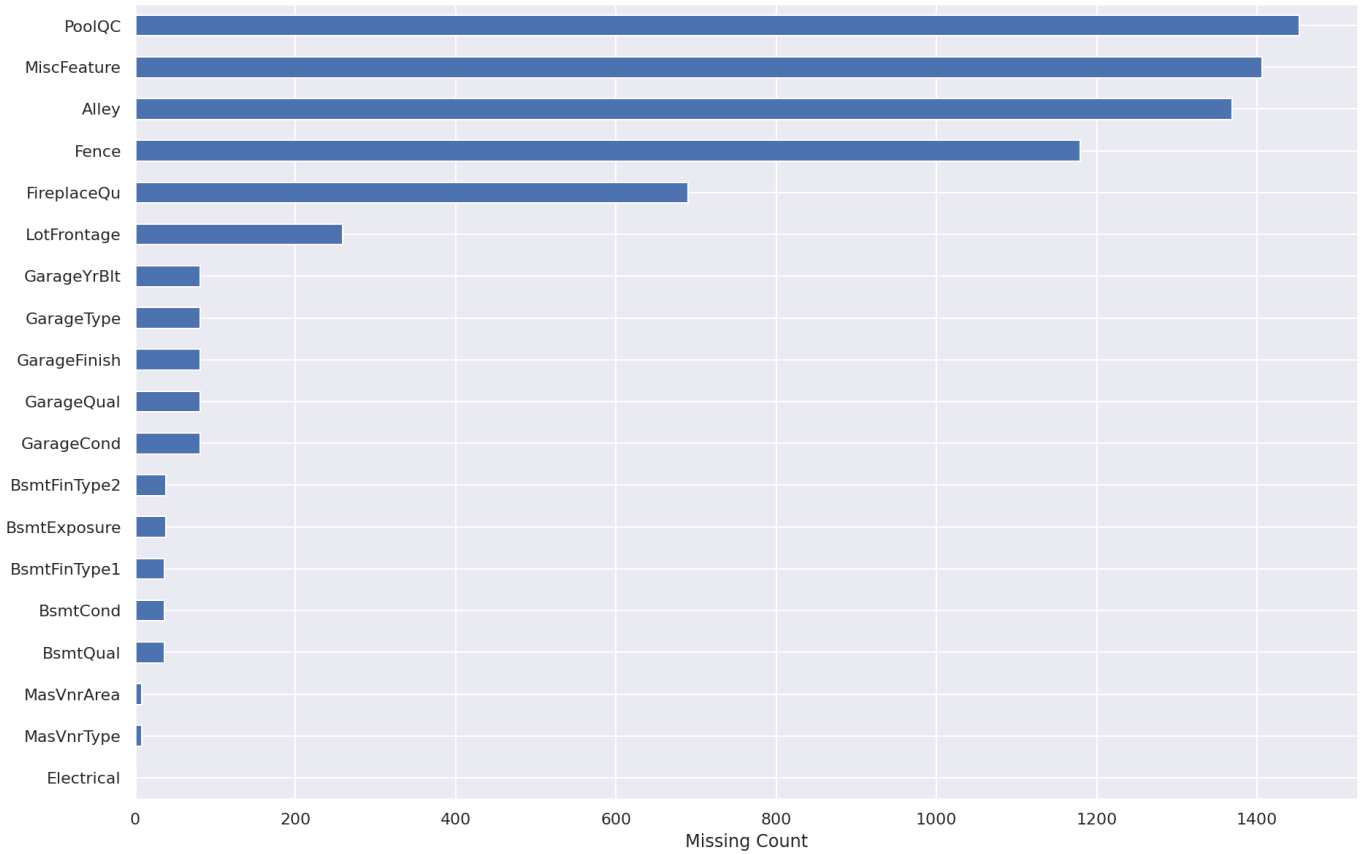
→ Nên vì mô hình được đánh giá theo RMSLE

count	1460.000000
mean	180921.195890
std	79442.502883
min	34900.000000
25%	129975.000000
50%	163000.000000
75%	214000.000000
max	755000.000000



Dữ liệu thiếu

	Column Name	Count	Ratio
0	LotFrontage	259	17.740000
1	Alley	1369	93.770000
2	MasVnrType	8	0.550000
3	MasVnrArea	8	0.550000
4	BsmtQual	37	2.530000
5	BsmtCond	37	2.530000
6	BsmtExposure	38	2.600000
7	BsmtFinType1	37	2.530000
8	BsmtFinType2	38	2.600000
9	Electrical	1	0.070000
10	FireplaceQu	690	47.260000
11	GarageType	81	5.550000
12	GarageYrBlt	81	5.550000
13	GarageFinish	81	5.550000
14	GarageQual	81	5.550000
15	GarageCond	81	5.550000
16	PoolQC	1453	99.520000
17	Fence	1179	80.750000
18	MiscFeature	1406	96.300000



Dữ liệu thiếu (cont)

- Hầu hết đều thiếu có chủ đích
- Một số thuộc tính thiếu thực sự (không đáng kể)
  - LotFrontage, MasVnrType, MasVnrArea, Electrical

Tiền xử lý

- Dữ liệu chỉ có 1 tập

→Tách tập validation để đánh giá mô hình (tỉ lệ 7:3)

- Fill dữ liệu thiếu (mean cho số, ‘Missing’ cho phân loại)
- Onehot cho các cột phân loại

Thử nghiệm mô hình

Mô hình

- Lasso
- Ridge
- Gradient Boosting
- XGBoost

Lasso Regression

- Least absolute shrinkage and selection operator
- Mô hình tuyến tính có sử dụng L1 regularize
- Hyperparameter:
  - alpha

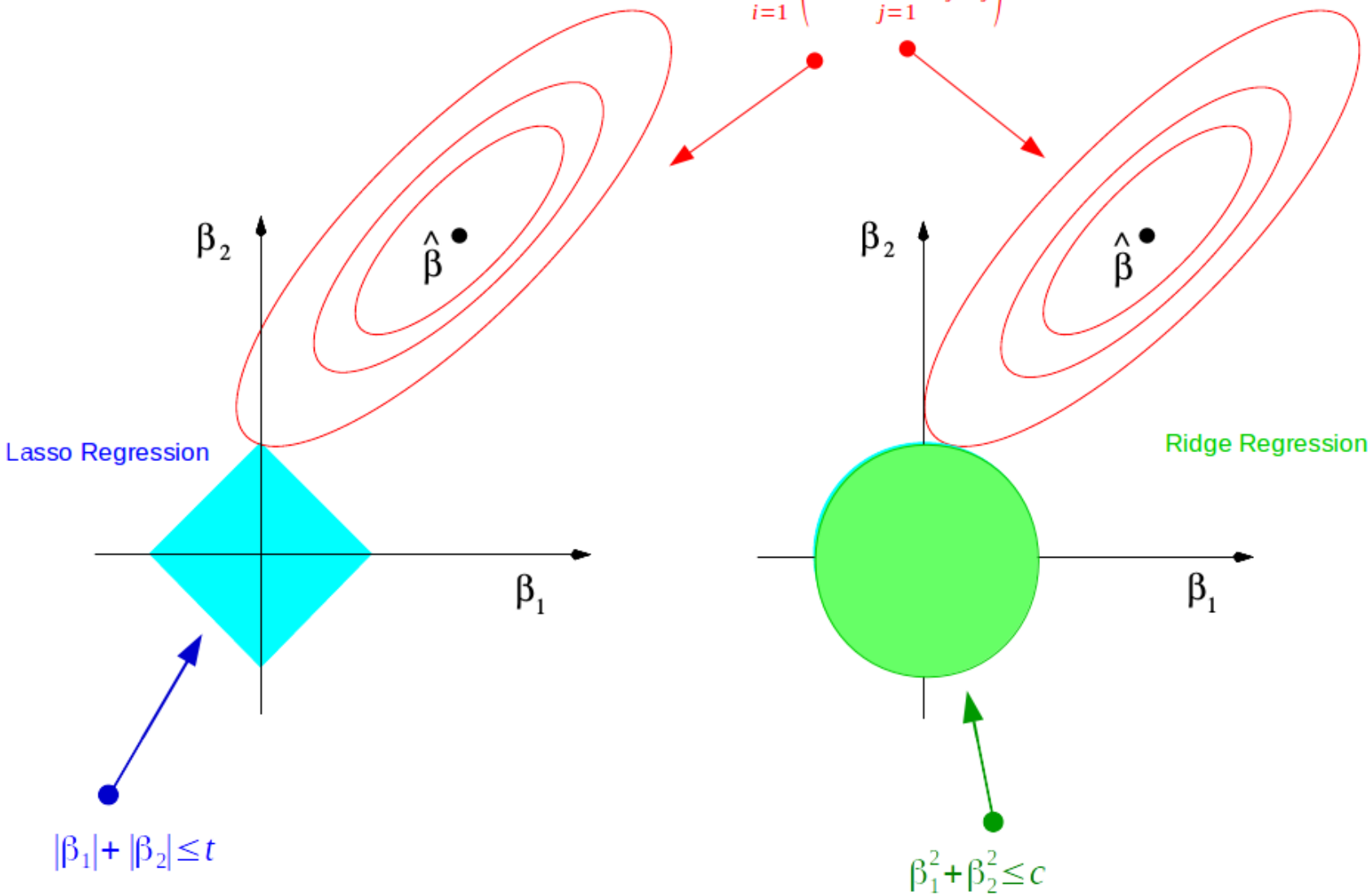
Ridge Regression

- Mô hình tuyến tính có sử dụng L2 regularize
- Hyperparameter:
  - alpha

Dimension Reduction of Feature Space with LASSO

Linear Regression Cost function

$$\sum_{i=1}^{M}\left(y_i-\sum_{j=1}^{j=2}\beta_j x_{ij}\right)^2$$



L2 hay L1?

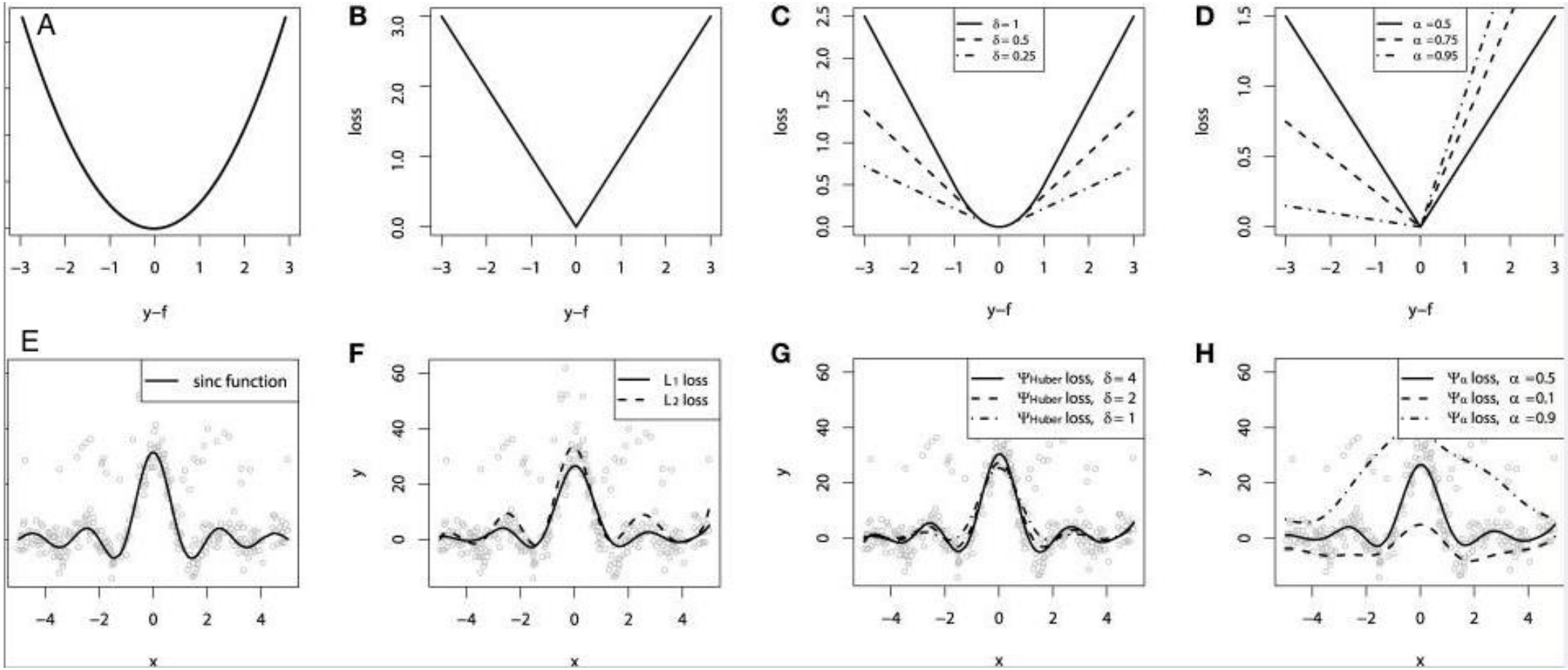
- Trong nhiều trường hợp, L1 regularize thường cho ra bộ trọng số sparse (tức có nhiều số 0)

→ Tránh được overfit tốt hơn

Gradient Boosting

- Hyperparameter:
  - loss: thử ls hoặc huber
  - alpha

So sánh các độ lỗi của GBR



So sánh các độ lỗi của GBR

- L2 là độ lỗi cơ bản, chênh lệch càng lớn độ lỗi càng lớn gấp bội, kết quả cho ra thường là mean
- L1 tuyến tính theo chênh lệch trong dự đoán, kết quả cho ra thường là median → Tránh được overfit
- Huber kết hợp L1, L2:
  - Chênh lệch lớn → L1
  - Chênh lệch nhỏ → L2

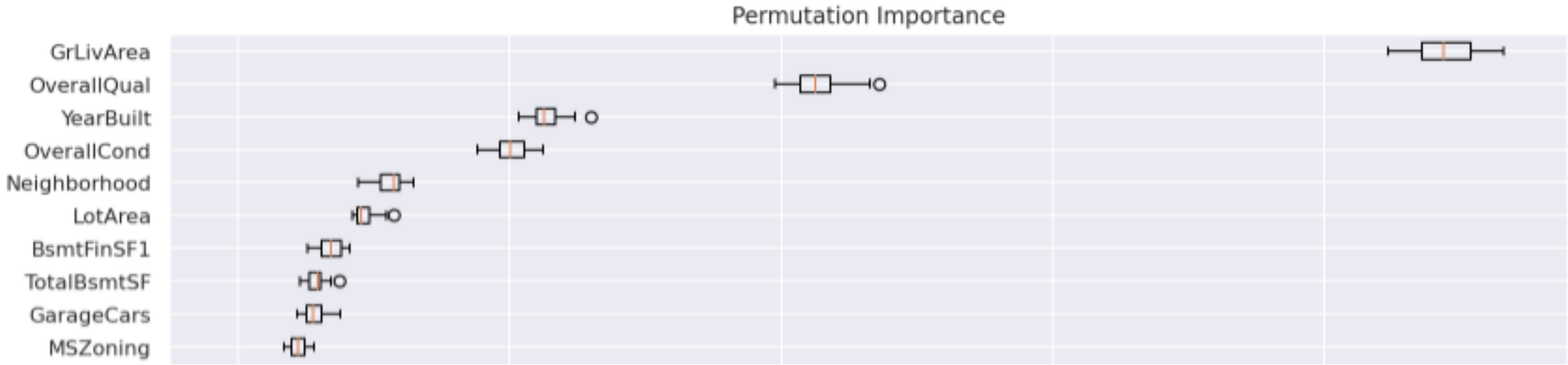
XGBoost

- Hyperparameter:
  - learning rate

Kết quả

Model	Test score	Best params
GBR	0.13300	loss=huber; alpha=0.8
Ridge	0.13253	alpha=0.05
Lasso	0.12856	alpha=0.0001
XGBoost	0.13641	lr=0.2
RidgeCV	0.13277	alpha=0.3
LassoCV	0. 12398	alpha=0.0005

Permutation importance



Ensemble

Voting (averaging)

- Kết hợp nhiều mô hình, kết quả dự đoán được lấy trung bình
- Kết hợp: gbr, ridge_cv, lasso_cv từ thử nghiệm trước

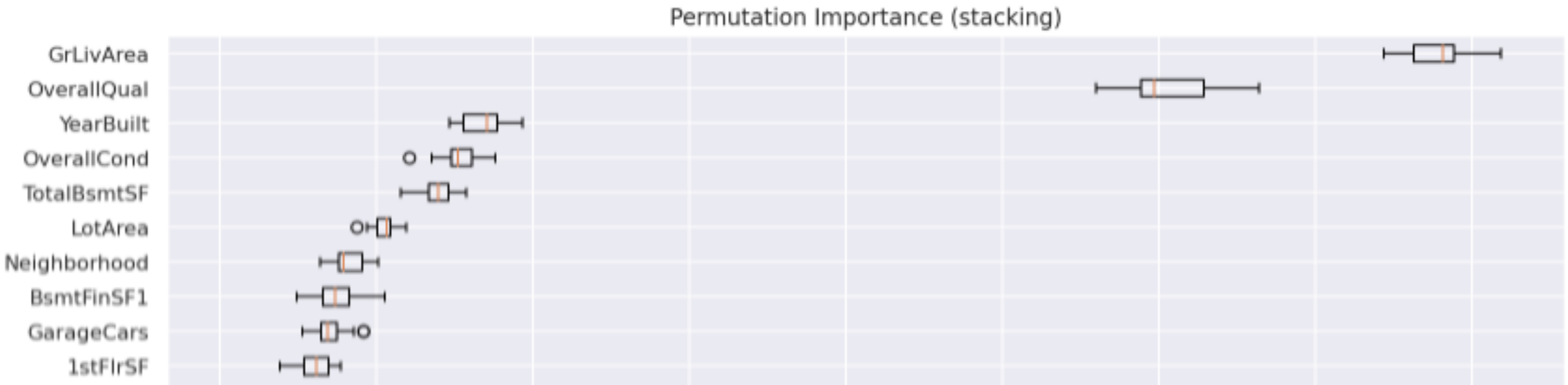
→ RSMLE: 0.12213

Stacking

- Kết hợp nhiều mô hình, kết quả dự đoán được chạy qua thêm một 1 mô hình
- Kết hợp: gbr, ridge_cv, lasso_cv từ thử nghiệm trước
- Mô hình đầu cuối: gbr với loss=lad (L1)

→ RSMLE: 0.11965

Permutation importance



Permutation importance

- GrLivArea: diện tích nhà (không tính hầm)
- OverallQual: chất lượng căn nhà: vật liệu, mức độ hoàn thiện (1 ~ 10)
- YearBuilt: thời điểm xây lần đầu
- OverallCond: đánh giá tổng quan tình trạng căn nhà (1 ~ 10)
- Neighborhood: gần với địa danh nào của thành phố (trường học, đường, cửa hàng,...)
- LotArea: diện tích lô đất
- BsmtFinType1: diện tích hầm loại 1
- TotalBsmtSF: tổng diện tích hầm
- GarageCars: sức chứa (số xe) của garage
- MSZoning: mục đích sử dụng chung được đăng ký với chính quyền (nông nghiệp, công nghiệp, kinh doanh,...)

Một số thử nghiệm thêm

Link drive: shorturl.at/eELS6

- Ver 1: gốc
- Ver 2: bỏ log transform → kết quả tệ đi rất nhiều
- Ver 3: tiền xử lý sơ xài → kết quả tệ 1 chút (~0.001 RMSLE)
- Ver 4: Onehot cho cột ordinal → kết quả tệ rất nhiều
- Ver 5: tạo feature mới bằng cách cộng một số feature cũ → kết quả giữ nguyên
- Ver 6: sử dụng Robust Scaler → kết quả giữ nguyên
- Ver 7: giữ feature trước và sau log → Kết quả tệ đi
- Ver 8: biến đổi một số cột có dạng ordinal thành số → Kết quả tệ đi

Tổng kết

Lời cảm ơn

Tham khảo

[1]. House Prices - Advanced Regression Techniques | Kaggle

[2]. House Price Prediction with Creative Feature Engineering and Advanced Regression Techniques | Data Science Blog (nycdatascience.com)

[3]. #1 House Prices Solution [top 1%] | Kaggle

[4]. Feature Engineering for House Prices | Kaggle

[5]. Ridge and Lasso Regression: L1 and L2 Regularization | by Saptashwa Bhattacharyya | Towards Data Science

[6]. 5 Regression Loss Functions All Machine Learners Should Know | by Prince Grover | Heartbeat (fritz.ai)