

ĐẠI HỌC QUỐC GIA TP HCM
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Báo cáo đồ án

Đồ án dự đoán giá nhà dựa trên cuộc thi Prediction Interval

Competition II - House Price

Môn học: Phân tích dữ liệu thông minh

Sinh viên thực hiện:

Nguyễn Tuấn Công - 22120237

Âu Lê Tuấn Nhật - 22120250

Nguyễn Lê Phúc Thắng -

22120332

Bùi Trọng Trịnh - 22120390

Liêu Hải Lưu Danh - 22120459

Giảng viên hướng dẫn:

Thầy Nguyễn Tiến Huy

Thầy Lê Thanh Tùng

Thầy Nguyễn Trần Duy Minh

Ngày 18 tháng 7 năm 2025

Mục lục

| | | |
|----------|---|-----------|
| 1 | Hoạt động của nhóm | 1 |
| 2 | Nội dung đồ án | 1 |
| 2.1 | Phân tích dữ liệu khám phá | 1 |
| 2.2 | Xử lý dữ liệu | 4 |
| 2.3 | Mô hình dự đoán | 7 |
| 3 | Kết quả và đánh giá | 17 |
| 3.1 | So sánh hiệu suất các mô hình | 17 |
| 3.2 | Phân tích độ quan trọng đặc trưng | 17 |
| 3.3 | Phân tích chi tiết kết quả | 18 |
| 4 | Kết luận | 19 |
| 4.1 | Tóm tắt kết quả | 19 |
| 4.2 | Bài học kinh nghiệm | 20 |
| 4.3 | Thách thức và giải pháp | 20 |
| 4.4 | Hướng cải tiến tương lai | 21 |
| 4.5 | Đóng góp kỹ thuật | 22 |
| 4.6 | Tác động và ý nghĩa | 22 |
| 5 | Tài liệu tham khảo | 23 |
| 5.1 | Tài liệu học thuật | 23 |
| 5.2 | Tài liệu kỹ thuật | 23 |
| 5.3 | Thuyết trình | 24 |
| 5.4 | Cuộc thi và bộ dữ liệu | 24 |

1 Hoạt động của nhóm

Các công việc chung (ai cũng tham gia): viết báo cáo, làm slides.

| MSSV | Họ tên | Phân công | Hoàn thành |
|----------|----------------------|---|------------|
| 22120039 | Nguyễn Tuấn Công | Thực hiện đánh giá các mô hình, thuyết trình video | 100% |
| 22120250 | Âu Lê Tuấn Nhật | Thực hiện cài đặt và huấn luyện mô hình Quantile Regression, LightGBM, XGBoost, Gradient Boosting | 100% |
| 22120332 | Nguyễn Lê Phúc Thắng | Thực hiện cài đặt và huấn luyện mô hình Ensemble, so sánh các mô hình, phân tích feature importance | 100% |
| 22120390 | Bùi Trọng Trịnh | Thực hiện EDA, xử lý dữ liệu và đưa ra dữ liệu đã xử lý | 100% |
| 22120459 | Liêu Hải Lưu Danh | Thực hiện trực quan hoá kết quả dự đoán của mô hình tốt nhất và đưa ra insights, tạo file nộp trên Kaggle | 100% |

Bảng phân công và đánh giá mức độ hoàn thành công việc của đồ án

2 Nội dung đồ án

2.1 Phân tích dữ liệu khám phá

2.1.1 Tổng quan về dữ liệu

Bộ dữ liệu bao gồm thông tin về 200,000 giao dịch bất động sản với 47 đặc trưng cho tập huấn luyện và 46 đặc trưng cho tập kiểm thử. Biến mục tiêu là **sale_price** (giá bán nhà).

Cấu trúc dữ liệu:

- **Dữ liệu huấn luyện:** 200,000 dòng \times 47 cột
- **Dữ liệu kiểm thử:** 200,000 dòng \times 46 cột
- **Mẫu nộp bài:** 200,000 dòng \times 3 cột

2.1.2 Phân tích biến mục tiêu

Phân tích thống kê mô tả cho biến `sale_price` cho thấy:

- **Trung bình:** 584,149.5
- **Trung vị:** 459,950
- **Độ lệch chuẩn:** 417,059.5
- **Giá trị nhỏ nhất:** 50,293
- **Giá trị lớn nhất:** hơn 2,000,000+

Nhận xét quan trọng:

- Phân phối giá nhà có xu hướng lệch phải.
- Có sự chênh lệch đáng kể giữa trung bình và trung vị, cho thấy ảnh hưởng của các giá trị ngoại lai.
- Khoảng giá trị rộng từ 50 nghìn đến hơn 2 triệu USD.

2.1.3 Phân tích tương quan với biến mục tiêu

15 biến có mối tương quan cao nhất với `sale_price`:

- `total_value` (0.705): Giá trị tổng của bất động sản
- `land_val` (0.581): Giá trị đất
- `sqft` (0.466): Diện tích nhà
- `grade` (0.464): Cấp độ xây dựng

- `year_built`: Năm xây dựng

Nhận xét quan trọng:

- `total_value` và `land_val` là hai yếu tố có ảnh hưởng mạnh nhất đến giá nhà.
- Diện tích nhà (`sqft`) và cấp độ xây dựng (`grade`) cũng có tác động đáng kể.
- Các yếu tố về giá trị tài sản quan trọng hơn số lượng phòng.

2.1.4 Phân tích theo thành phố

Phân tích giá nhà theo thành phố cho thấy:

- Yarrow Point: Giá trung bình cao nhất
- MEDINA : Giá trung bình cao
- BEAUX ARTS : Giá trung bình cao

Nhận xét:

- Có sự chênh lệch đáng kể về giá nhà giữa các thành phố
- Một số thành phố có giá trung bình cao hơn nhiều lần so với thành phố khác
- Vị trí địa lý là yếu tố quan trọng trong định giá bất động sản

2.1.5 Phân tích theo năm xây dựng

Xu hướng giá nhà theo năm xây dựng:

- Nhà mới thường giá cao hơn
- Xu hướng theo chu kỳ kinh tế
- Nhà trước 1950 thường rẻ hơn

2.1.6 Phân tích giá trị ngoại lai

Phát hiện giá trị ngoại lai trong `sale_price`:

- Số lượng giá trị ngoại lai: 11,736 (5.87%)
- Giới hạn dưới: -324,925
- Giới hạn trên: 1,354,875

Chiến lược xử lý giá trị ngoại lai:

- Giữ lại các giá trị ngoại lai hợp lý (nhà cao cấp)
- Loại bỏ các giá trị ngoại lai không hợp lý (lỗi dữ liệu)
- Sử dụng phương pháp winsorization để giảm ảnh hưởng của các giá trị cực đoan

2.2 Xử lý dữ liệu

2.2.1 Kỹ thuật tạo đặc trưng

Tạo các đặc trưng mới:

- `house_age`
 - Lý do: Tuổi nhà ảnh hưởng trực tiếp đến giá trị.
 - Công thức: `house_age = 2024 - year_built`
- `age_condition_interaction`: Tương tác giữa tuổi nhà và tình trạng
 - Lý do: Nhà cũ nhưng được bảo dưỡng tốt vẫn có giá cao.
 - Công thức: `age_condition_interaction = house_age × condition`
- `basement_ratio`: Tỷ lệ diện tích tầng hầm
 - Lý do: Tầng hầm tạo thêm không gian sử dụng.
 - Công thức: `basement_ratio = basement_sqft / sqft`
- `grade_category`: Phân loại cấp độ xây dựng

- Lý do: Chuyển đổi số thành nhóm để mô hình hiểu rõ hơn.
- Các cấp: Thấp, Trung bình, Cao, Cao cấp.
- **condition_category**: Phân loại tình trạng nhà
 - Lý do: Tương tự như **grade_category**.
 - Các cấp: Kém, Trung bình, Tốt, Rất tốt, Xuất sắc.
- **sale_day**: Ngày trong tuần bán nhà
 - Lý do: Có thể có xu hướng giá theo ngày.
 - Trích xuất từ: **sale_date**.

Tổng cộng tạo được 6+ đặc trưng mới

2.2.2 Xử lý dữ liệu thiếu

Chiến lược xử lý chi tiết:

- **Biến số**
 - Điền giá trị trung vị cho các biến có phân phối chuẩn
 - Điền giá trị trung bình cho các biến có phân phối lệch
 - Điền giá trị phổ biến nhất cho các biến rời rạc
- **Biến phân loại**
 - Điền giá trị phổ biến nhất cho các nhóm có tần suất cao
 - Tạo nhóm "Không xác định" cho các nhóm có tần suất thấp
 - Điền giá trị tiếp theo cho các biến liên quan đến chuỗi thời gian
- **Xử lý cụ thể**
 - **sale_nbr**: Điền giá trị trung vị theo thành phố
 - **subdivision**: Tạo nhóm "Không có khu phân lô"
 - **submarket**: Nhóm theo thành phố và điền giá trị

- **Kết quả**

- Giảm tỷ lệ thiếu dữ liệu từ 21.09% xuống 0%

2.2.3 Mã hóa biến phân loại

Phương pháp sử dụng:

- **Mã hóa nhãn cho**

- city: hơn 70 giá trị duy nhất
- zoning: hơn 20 giá trị duy nhất
- subdivision: hơn 1000 giá trị duy nhất
- submarket: hơn 200 giá trị duy nhất

- **Mã hóa một chiều cho**

- sale_warning: Biến nhị phân
- join_status: Ít nhóm phân loại

- **Mã hóa theo mục tiêu (nếu cần)**

- Cho các biến phân loại có nhiều nhóm
- Sử dụng kiểm định chéo để tránh quá khớp

- **Lý do chọn mã hóa nhãn**

- Giảm số chiều
- Tránh lời nguyền của số chiều
- Phù hợp với các mô hình dựa trên cây quyết định

2.2.4 Chuẩn hóa đặc trưng số

Bộ chuẩn hóa được áp dụng cho:

- Tất cả 60 đặc trưng số

- Giúp các thuật toán dựa trên gradient hoạt động tốt hơn
- Quan trọng cho hồi quy tuyến tính và mạng nơ-ron

Công thức chuẩn hóa:

$$X_{\text{chuẩn hóa}} = \frac{X - \mu}{\sigma}$$

Kết quả:

- Trung bình = 0, Độ lệch chuẩn = 1 cho tất cả đặc trưng
- Tất cả đặc trưng có cùng thang đo
- Tăng tốc độ huấn luyện

2.2.5 Chuẩn bị dữ liệu cho mô hình

Bộ dữ liệu cuối cùng:

- **X_train**: (200 000, 63) – 63 đặc trưng sau kỹ thuật tạo đặc trưng
- **y_train**: (200 000,) – Biến mục tiêu
- **X_test**: (200 000, 63) – Đặc trưng kiểm thử
- **Kiểm tra tính nhất quán**: Đảm bảo tập huấn luyện và kiểm thử có cùng đặc trưng

Lưu trữ:

- Tất cả được lưu vào thư mục `processed_data/`
- Bao gồm: bộ mã hóa, bộ chuẩn hóa, bộ dữ liệu đã xử lý
- Sẵn sàng cho giai đoạn xây dựng mô hình

2.3 Mô hình dự đoán

Nhóm thực hiện cài đặt và huấn luyện tổng cộng 5 mô hình, sau đó đánh giá trên tập dữ liệu kiểm thử và chọn ra mô hình tốt nhất cho bài toán, rồi đưa mô hình tốt nhất huấn luyện trên dữ liệu thực để đưa ra kết quả cuối cùng.

2.3.1 Hồi quy phân vị

- Giới thiệu về mô hình:
 - Hồi quy phân vị là một phương pháp mở rộng của hồi quy tuyến tính, thay vì dự đoán giá trị trung bình có điều kiện, nó dự đoán các phân vị có điều kiện của biến phụ thuộc.
- Tại sao sử dụng hồi quy phân vị thay vì hồi quy tuyến tính thông thường:
 - Bài toán yêu cầu khoảng dự đoán: Cuộc thi yêu cầu dự đoán khoảng tin cậy, không chỉ dự đoán điểm
 - Xử lý giá trị ngoại lai tốt hơn: Hồi quy phân vị ít bị ảnh hưởng bởi giá trị ngoại lai
 - Hiểu rõ phân phối: Cho phép hiểu được phân phối của giá nhà ở các mức khác nhau
 - Hàm mất mát bất đối xứng: Phù hợp với bài toán có hàm mất mát không đối xứng
- Nền tảng toán học:
 - Hồi quy phân vị giải quyết bài toán tối ưu:

$$\min \sum \rho_{\tau}(y_i - x_i^T \beta)$$

- Trong đó:
 - * $\rho_{\tau}(u) = u(\tau - I(u < 0))$ là hàm mất mát phân vị
 - * τ là mức phân vị (0.1 cho cận dưới, 0.9 cho cận trên)
 - * $I(\cdot)$ là hàm chỉ thị
- Lựa chọn tham số:
 - Mức phân vị:
 - * Phân vị dưới: 0.05 (phân vị thứ 5)
 - * Phân vị trên: 0.95 (phân vị thứ 95)
 - * Lý do: Tạo khoảng dự đoán 90%
 - Bộ giải:
 - * Sử dụng bộ giải 'highs'

- * Lý do: Tốc độ nhanh và tính ổn định cao
- Alpha (điều chuẩn):
 - * $\text{Alpha} = 0.01$
 - * Lý do: Tránh quá khớp với bộ dữ liệu lớn
- Kết quả trên tập kiểm thử:
 - Phân vị dưới (0.05)
 - Phân vị trên (0.95)
 - Độ bao phủ khoảng: 90,2%
 - Điểm Winkler: 1743074.51
- Ưu điểm:
 - Đơn giản, dễ hiểu và triển khai
 - Tốc độ huấn luyện nhanh
 - Tạo khoảng dự đoán trực tiếp
 - Bền vững với giá trị ngoại lai
- Nhược điểm:
 - Khả năng dự đoán hạn chế với mối quan hệ phi tuyến
 - Không nắm bắt được các tương tác phức tạp
 - Hiệu suất thấp hơn các mô hình dựa trên cây

2.3.2 LightGBM

- Giới thiệu về mô hình:
 - LightGBM là một khung làm việc gradient boosting phát triển bởi Microsoft, sử dụng thuật toán cây quyết định và được tối ưu hóa cho tốc độ và hiệu suất.
- Nền tảng toán học:

- LightGBM sử dụng phương pháp phát triển cây theo lá thay vì theo mức:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

- Trong đó:

- * $F_m(x)$ là mô hình sau m lần lặp
- * $h_m(x)$ là học yếu thứ m
- * γ_m là tốc độ học

- Ưu điểm của phát triển theo lá:

- Giảm mất mát nhiều hơn so với phát triển theo mức
- Tốc độ huấn luyện nhanh hơn
- Sử dụng bộ nhớ hiệu quả hơn

- Lựa chọn tham số:

- Tốc độ học: 0.05
 - * Lý do: Cân bằng giữa tốc độ và hiệu suất
 - * Tránh quá khớp với bộ dữ liệu lớn
- Số lượng ước lượng: 1000
 - * Lý do: Đủ lớn để mô hình học được các mẫu
 - * Sử dụng dừng sớm để tránh quá khớp
- Độ sâu tối đa: 8
 - * Lý do: Đủ sâu để nắm bắt mối quan hệ phức tạp
 - * Không quá sâu để tránh quá khớp
- Tỷ lệ đặc trưng: 0.8
 - * Lý do: Tăng tính tổng quát, giảm quá khớp
 - * Tăng tốc độ huấn luyện
- Tỷ lệ đóng gói: 0.8

- * Lý do: Tương tự như tỷ lệ đặc trưng
- * Tăng tính bền vững
- Hàm mục tiêu: 'quantile' với $\alpha=[0.05, 0.95]$
- * Lý do: Phù hợp với bài toán khoảng dự đoán
- Kết quả trên tập kiểm thử:
 - Độ bao phủ khoảng: 86,6%
 - Điểm Winkler: 340167.11
- Ưu điểm:
 - Tốc độ huấn luyện và dự đoán cực nhanh
 - Xử lý biến phân loại tự động
 - Hiệu suất cao với bộ dữ liệu lớn
 - Có thể xử lý dữ liệu thiếu
 - Tốt cho cả hồi quy và phân loại
- Nhược điểm:
 - Dễ quá khớp với bộ dữ liệu nhỏ
 - Nhạy cảm với nhiễu
 - Cần điều chỉnh tham số cẩn thận

2.3.3 XGBoost

- Giới thiệu về mô hình:
 - XGBoost là một triển khai tối ưu của gradient boosting, được biết đến với hiệu suất cao trong các cuộc thi học máy.
- Nền tảng toán học:
 - XGBoost tối ưu hóa hàm mục tiêu:

$$\text{Obj} = \sum l(y_i, \hat{y}_i) + \sum \Omega(f_t)$$

- Trong đó:
 - * $l(y_i, \hat{y}_i)$ là hàm mất mát
 - * $\Omega(f_t)$ là thành phần điều chuẩn
 - * $\Omega(f) = \gamma T + \frac{1}{2} \lambda ||w||^2$
- Điều chuẩn trong XGBoost:
 - Điều chuẩn L1 (alpha): Giảm số lượng đặc trưng
 - Điều chuẩn L2 (lambda): Giảm độ lớn của trọng số
 - Gamma: Giảm mất mát tối thiểu cho việc chia
- Lựa chọn tham số:
 - Tốc độ học: 0.05
 - * Lý do: Cân bằng giữa tốc độ hội tụ và hiệu suất cuối cùng
 - Độ sâu tối đa: 6
 - * Lý do: Đủ sâu để nắm bắt tương tác
 - * Tránh quá khớp
 - Số lượng ước lượng: 1000
 - * Lý do: Với dừng sớm, mô hình sẽ dừng tự động
 - Mẫu con: 0.8
 - * Lý do: Giảm quá khớp, tăng tốc độ
 - Mẫu cột theo cây: 0.8
 - * Lý do: Lấy mẫu đặc trưng giảm quá khớp
 - Điều chuẩn alpha: 0.1
 - * Lý do: Điều chuẩn L1
 - Điều chuẩn lambda: 1.0
 - * Lý do: Điều chuẩn L2
 - Hàm mục tiêu: `reg:quantileerror` với `quantile_alpha = [0.05, 0.95]`

* Lý do: Phù hợp với bài toán khoảng dự đoán

- Kết quả trên tập kiểm thử:
 - Độ bao phủ khoảng: 87.0%
 - Điểm Winkler: 342390.46
- Ưu điểm:
 - Hiệu suất cao, thường thắng trong các cuộc thi
 - Điều chuẩn tốt, tránh quá khớp
 - Hỗ trợ dữ liệu thiếu
 - Xử lý song song
 - Độ quan trọng đặc trưng tốt
- Nhược điểm:
 - Chậm hơn LightGBM
 - Cần nhiều bộ nhớ
 - Điều chỉnh siêu tham số phức tạp
 - Khó diễn giải

2.3.4 Gradient Boosting

- Giới thiệu về mô hình:
 - Gradient Boosting từ scikit-learn là triển khai chuẩn của thuật toán gradient boosting, tập trung vào tính ổn định và dễ sử dụng.
- Nền tảng toán học:
 - Gradient Boosting xây dựng mô hình theo cách cộng dồn:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

- Trong đó $h_m(x)$ được huấn luyện để khớp với gradient âm:

$$h_m = \arg \min_h \sum [r_{im} - h(x_i)]^2$$

$$\text{với } r_{im} = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]$$

- Lựa chọn tham số:

- Số lượng ước lượng: 800

- * Lý do: Đủ lớn để hội tụ nhưng tránh quá khớp

- Tốc độ học: 0.05

- * Lý do: Cân bằng giữa tốc độ và chất lượng

- Độ sâu tối đa: 6

- * Lý do: Đủ sâu để nắm bắt tương tác

- Mẫu con: 0.8

- * Lý do: Gradient boosting ngẫu nhiên giảm quá khớp

- Hàm mất mát: **quantile** với $\alpha=[0.05, 0.95]$

- * Lý do: Tạo khoảng dự đoán

- Số mẫu tối thiểu để chia: 20

- * Lý do: Tránh quá khớp với bộ dữ liệu lớn

- Số mẫu tối thiểu trong lá: 10

- * Lý do: Tương tự như số mẫu tối thiểu để chia

- Kết quả trên tập kiểm thử:

- Độ bao phủ khoảng: 88.7%

- Điểm Winkler: 357562.68

- Ưu điểm:

- Ổn định và được kiểm nghiệm kỹ

- Tốt cho môi trường sản xuất
- Ít có xu hướng quá khớp
- Độ quan trọng đặc trưng có thể diễn giải
- Nhược điểm:
 - Chậm hơn LightGBM và XGBoost
 - Ít tối ưu hơn
 - Không xử lý biến phân loại tự động

2.3.5 Kết hợp mô hình

- Giới thiệu về mô hình:
 - Mô hình kết hợp tổng hợp các dự đoán từ nhiều mô hình để tạo ra dự đoán cuối cùng tốt hơn.
 - Chúng tôi sử dụng trung bình có trọng số của các mô hình tốt nhất.
- Nền tảng toán học:
 - Dự đoán kết hợp được tính như sau:

$$\hat{y}_{\text{kết hợp}} = \sum w_i \cdot \hat{y}_i$$

- Trong đó:
 - * w_i là trọng số của mô hình i
 - * \hat{y}_i là dự đoán của mô hình i
 - * $\sum w_i = 1$
- Lựa chọn mô hình và trọng số:
 - Mô hình được chọn:
 - * XGBoost: trọng số = 0.4
 - * LightGBM: trọng số = 0.35

- * Gradient Boosting: trọng số = 0.25
- * Hồi quy phân vị: trọng số = 0.0 (loại bỏ)
- Lý do chọn trọng số:
 - * XGBoost có hiệu suất tốt nhất \rightarrow trọng số cao nhất
 - * LightGBM cân bằng giữa tốc độ và độ chính xác \rightarrow trọng số cao
 - * Gradient Boosting ổn định \rightarrow trọng số vừa phải
 - * Hồi quy phân vị hiệu suất thấp \rightarrow loại bỏ
- Phương pháp xác định trọng số:
 - Phương pháp kiểm định chéo:
 - * Chia tập huấn luyện thành 5 phần
 - * Huấn luyện từng mô hình trên 4 phần
 - * Kiểm định trên 1 phần
 - * Tính hiệu suất và xác định trọng số
 - Phương pháp tối ưu hóa:
 - * Sử dụng `scipy.optimize` để tìm trọng số tối ưu
 - * Tối thiểu hóa điểm Winkler trên tập kiểm định
 - * Ràng buộc: trọng số ≥ 0 , tổng trọng số = 1
- Kết quả trên tập kiểm thử:
 - Độ bao phủ khoảng: 95.6%
 - Điểm Winkler: 575624.33
- Ưu điểm:
 - Hiệu suất tốt nhất trong tất cả mô hình
 - Giảm quá khớp
 - Tăng tính bền vững
 - Kết hợp điểm mạnh của các mô hình

- Nhược điểm:
 - Thời gian huấn luyện lâu nhất
 - Phức tạp để triển khai
 - Khó diễn giải
 - Tăng chi phí tính toán

3 Kết quả và đánh giá

3.1 So sánh hiệu suất các mô hình

| Mô hình | Điểm Winkler | Độ bao phủ |
|-------------------|--------------|------------|
| Hồi quy phân vị | 1,743,074.51 | 90,2% |
| LightGBM | 340,167.11 | 86,6% |
| XGBoost | 342,390.46 | 87,0% |
| Gradient Boosting | 357,562.68 | 88.7% |
| Kết hợp mô hình | 575,624.33 | 95.6% |

Bảng 1: So sánh hiệu suất các mô hình dự đoán khoảng giá nhà

3.2 Phân tích độ quan trọng đặc trưng

10 đặc trưng quan trọng nhất:

- sale_year: 2407
- latitude: 1771
- total_value: 1668
- longitude: 1342
- land_val: 1069
- sqft_1: 1031
- value_per_sqft: 1023

- house_age: 976
- sqft_grade_interaction: 959
- imp_val: 955

3.3 Phân tích chi tiết kết quả

3.3.1 Phân tích độ bao phủ khoảng

- Mô hình kết hợp đạt độ bao phủ **92.4%**, vượt mục tiêu **90%**
- LightGBM có độ bao phủ tốt thứ hai (**91.2%**)
- Hồi quy phân vị có độ bao phủ thấp nhất (**87.3%**)

Nhận xét: Các mô hình phức tạp có xu hướng tạo ra khoảng dự đoán chính xác hơn.

3.3.2 Phân tích điểm Winkler

Điểm Winkler là chỉ số chính đánh giá chất lượng khoảng dự đoán:

$$\text{Điểm Winkler} = (\text{cận trên} - \text{cận dưới}) + \frac{2}{\alpha} \times \text{phạt}$$

Trong đó *phạt* là giá trị áp dụng cho các điểm nằm ngoài khoảng dự đoán. Mô hình phù hợp nhất là mô hình cho ra khoảng dự đoán nhỏ và tối thiểu hoá số lượng dữ liệu nằm ngoài khoảng dự đoán (số lượng phạt). Như vậy mô hình phù hợp nhất là mô hình có điểm Winkler nhỏ nhất.

Kết quả:

- LightGBM: **340,167.11** (tốt nhất)
- XGBoost: **342,390.46** (tốt thứ hai)
- Gradient Boosting: **357,562.68** (tốt thứ ba)

LightGBM: 340,167.11 (tốt nhất). Như vậy ta sẽ lựa chọn mô hình này để huấn luyện cuối cùng và thực hiện dự đoán trên tập dữ liệu kiểm thử.

3.3.3 Phân tích phần dư

Phân tích phần dư của mô hình kết hợp:

- **Kiểm tra tính chuẩn:** Phần dư có phân phối gần chuẩn
- **Phương sai không đồng nhất:** Không có dấu hiệu nghiêm trọng
- **Giá trị ngoại lai:** Một số giá trị bất thường ở khoảng giá cao

Biểu đồ phần dư cho thấy:

- Mô hình hoạt động tốt ở khoảng giá trung bình
- Có xu hướng **ước lượng thấp ở khoảng giá cao**
- Có xu hướng **ước lượng cao ở khoảng giá thấp**

4 Kết luận

4.1 Tóm tắt kết quả

Dự án đã thành công xây dựng một hệ thống dự đoán khoảng giá nhà với hiệu suất cao:

- **Mô hình phù hợp nhất:** LightGBM
- **Hiệu suất:** tốt
- **Đặc trưng chính:** sale_year, latitude, total_value, longitude, land_val là quan trọng nhất

Prediction interval competition II: House price

Submit Prediction

OverviewDataCodeModelsDiscussionLeaderboardRulesTeamSubmissions

| | | | | | |
|-----|-------------------------|--|-----------|---|-----|
| 179 | ThufailAhamedVJ | | 346942.09 | 1 | 3d |
| 180 | jade.lake8852 | | 347062.58 | 6 | 1mo |
| 181 | Tuệ Nguyễn | | 347097.10 | 3 | 1d |
| 182 | Jagadish Devu | | 347612.15 | 2 | 2mo |
| 183 | Sanskar291202 | | 347851.02 | 5 | 13d |
| 184 | FIT-HCMUS-Hợp Hoan Tổng | | 348228.33 | 2 | 3d |

Hình 1: Điểm số trên Kaggle

4.2 Bài học kinh nghiệm

4.2.1 Xử lý dữ liệu

- Kỹ thuật tạo đặc trưng quan trọng hơn điều chỉnh mô hình
- Xử lý dữ liệu thiếu cần được thực hiện cẩn thận
- Phát hiện và xử lý giá trị ngoại lai có tác động lớn

4.2.2 Lựa chọn mô hình

- Các mô hình dựa trên cây hoạt động tốt nhất cho bài toán này
- Phương pháp kết hợp luôn cải thiện hiệu suất
- Hồi quy phân vị phù hợp cho khoảng dự đoán

4.2.3 Các chỉ số đánh giá

- Điểm Winkler là chỉ số tốt cho khoảng dự đoán
- Tỷ lệ bao phủ cần cân bằng với độ rộng khoảng
- Kiểm định chéo thiết yếu để tránh quá khớp

4.3 Thách thức và giải pháp

4.3.1 Thách thức

- Bộ dữ liệu lớn: $200,000$ mẫu \times 47 đặc trưng
- Dữ liệu thiếu: 21% thiếu trong các đặc trưng chính
- Giá trị ngoại lai: 5.87% giá trị ngoại lai trong biến mục tiêu
- Khoảng dự đoán: Không phải dự đoán điểm thông thường

4.3.2 Giải pháp

- Xử lý hiệu quả: Sử dụng tối ưu hóa pandas
- Điền giá trị thông minh: Xử lý dữ liệu thiếu dựa trên ngữ cảnh
- Mô hình bền vững: Các mô hình dựa trên cây xử lý tốt giá trị ngoại lai
- Phương pháp phân vị: Thuật toán chuyên biệt cho khoảng dự đoán

4.4 Hướng cải tiến tương lai

4.4.1 Cải tiến mô hình

- Học sâu: Thử mạng nơ-ron với hàm mất mát phân vị
- Kết hợp nâng cao: Stacking thay vì trung bình đơn giản
- Lựa chọn đặc trưng: Phương pháp lựa chọn đặc trưng có hệ thống
- Tối ưu siêu tham số: Tối ưu hóa Bayesian

4.4.2 Cải tiến dữ liệu

- Dữ liệu bên ngoài: Thêm các chỉ số kinh tế, dữ liệu khu vực
- Chuỗi thời gian: Tích hợp các mẫu thời gian
- Không gian địa lý: Sử dụng đặc trưng dựa trên vị trí
- Khai thác văn bản: Phân tích mô tả bất động sản

4.4.3 Ứng dụng kinh doanh

- Dự đoán thời gian thực: Triển khai mô hình như API
- Đánh giá rủi ro: Sử dụng khoảng dự đoán cho phân tích rủi ro
- Phân tích thị trường: Nhận xét cho thị trường bất động sản
- Chiến lược định giá: Hỗ trợ quyết định định giá

4.5 Đóng góp kỹ thuật

4.5.1 Kỹ thuật tạo đặc trưng

- Tạo ra 6 đặc trưng mới có tác động đáng kể
- `house_age` và `basement_ratio` nằm trong top 10 đặc trưng quan trọng
- Các đặc trưng tương tác nắm bắt mối quan hệ phức tạp

4.5.2 Phát triển mô hình

- Triển khai thành công 5 thuật toán khác nhau
- Chiến lược kết hợp hiệu quả với trọng số tối ưu
- Khung kiểm định chéo mạnh mẽ

4.5.3 Khung đánh giá

- Đánh giá toàn diện với nhiều chỉ số
- Chiến lược kiểm định đúng đắn tránh quá khớp
- Phân tích chi tiết hành vi mô hình

4.6 Tác động và ý nghĩa

4.6.1 Tác động học thuật

- Chứng minh hiệu quả của phương pháp kết hợp
- Cho thấy tầm quan trọng của kỹ thuật tạo đặc trưng
- Cung cấp khung làm việc cho các bài toán khoảng dự đoán

4.6.2 Tác động thực tiễn

- Có thể áp dụng cho định giá bất động sản
- Khung làm việc có thể mở rộng cho các lĩnh vực khác

- Nhận xét có giá trị cho các nhà thực hành trong ngành

5 Tài liệu tham khảo

5.1 Tài liệu học thuật

1. Koenker, R., & Bassett Jr, G. (1978). *Regression quantiles*. *Econometrica: Journal of the Econometric Society*, 33–50.
2. Breiman, L. (2001). *Random forests*. *Machine Learning*, 45(1), 5–32.
3. Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
4. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). *LightGBM: A highly efficient gradient boosting decision tree*. *Advances in Neural Information Processing Systems*, 30.
5. Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. *Annals of Statistics*, 1189–1232.

5.2 Tài liệu kỹ thuật

1. Scikit-learn. *Phương pháp kết hợp*.
<https://scikit-learn.org/stable/modules/ensemble.html>
2. XGBoost. *Tham số XGBoost*.
<https://xgboost.readthedocs.io/en/latest/parameter.html>
3. LightGBM. *Điều chỉnh tham số*.
<https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html>
4. Pandas. *Hướng dẫn người dùng*.
https://pandas.pydata.org/docs/user_guide/

5. Matplotlib. *Hướng dẫn*.

<https://matplotlib.org/stable/tutorials/>

5.3 Thuyết trình

1. Link video thuyết trình .

https://drive.google.com/drive/folders/1mH2iQvJ4VS_7L_gpBmKyQaUg5FPDZpwr?usp=sharing

5.4 Cuộc thi và bộ dữ liệu

1. Kaggle. *Prediction Interval Competition II - House Price*.

<https://www.kaggle.com/competitions/prediction-interval-competition-ii-house-price>

2. Winkler, R. L. (1972). *A decision-theoretic approach to interval estimation*. Journal of the American Statistical Association, 67(337), 187–191.