Phần 1: Related Work (Công trình liên quan)

1. **Các nghiên cứu liên quan trên Kaggle và các nguồn khác**

Tập dữ liệu Ames Housing đã được sử dụng trong nhiều nghiên cứu, bài viết, và notebook trên Kaggle, cũng như trong các nghiên cứu học thuật. Dưới đây là một số hướng nghiên cứu liên quan:

1. **Dự đoán giá nhà (House Price Prediction)**

* Cuộc thi trên Kaggle: Tập dữ liệu Ames Housing là nền tảng của cuộc thi "House Prices - Advanced Regression Techniques" trên Kaggle, nơi người tham gia xây dựng các mô hình để dự đoán SalePrice. Các phương pháp phổ biến bao gồm:
* Hồi quy tuyến tính (Linear Regression): Một số người dùng trên Kaggle đã áp dụng hồi quy tuyến tính cơ bản để xác định các yếu tố quan trọng ảnh hưởng đến giá nhà, như OverallQual (chất lượng tổng thể) và GrLivArea (diện tích tầng trệt). Một notebook trên Kaggle đã chỉ ra rằng OverallQual có tương quan 0.79 với SalePrice, là yếu tố dự đoán mạnh.
* Gradient Boosting: Các mô hình như XGBoost và LightGBM thường được sử dụng để đạt kết quả tốt hơn. Một số người tham gia đạt RMSE (Root Mean Squared Error) khoảng 0.11-0.12 trên tập kiểm tra bằng cách sử dụng XGBoost với các kỹ thuật xử lý giá trị thiếu và mã hóa biến phân loại.
* Mạng nơ-ron (Neural Networks): Một số nghiên cứu thử nghiệm mạng nơ-ron sâu, nhưng hiệu quả thường không vượt trội so với Gradient Boosting do tập dữ liệu có quy mô vừa phải.

1. **Phân tích dữ liệu khám phá (Exploratory Data Analysis - EDA)**

Nhiều notebook trên Kaggle tập trung vào EDA để hiểu rõ hơn về dữ liệu:

* Phân phối của SalePrice: Một phân tích phổ biến trên Kaggle cho thấy SalePrice có phân phối lệch phải (right-skewed). Người dùng thường áp dụng biến đổi logarit (log transformation) để làm cho phân phối gần với phân phối chuẩn hơn, giúp cải thiện hiệu suất của các mô hình hồi quy.
* Tương quan giữa các biến: Ma trận tương quan (correlation matrix) thường được vẽ để xác định các biến có mối quan hệ mạnh với SalePrice. Ngoài OverallQual và GrLivArea, các biến như TotalBsmtSF (diện tích tầng hầm) và GarageCars (số lượng xe có thể đỗ trong gara) cũng cho thấy tương quan cao.
* Xử lý giá trị ngoại lai (Outliers): Một số người dùng đã phát hiện các giá trị ngoại lai, chẳng hạn như các ngôi nhà có diện tích lớn (GrLivArea > 4000) nhưng giá bán thấp bất thường, và loại bỏ chúng để cải thiện mô hình.

1. **Kỹ thuật xử lý đặc trưng (Feature Engineering)**

Tập dữ liệu Ames Housing có 80 biến, nên kỹ thuật xử lý đặc trưng đóng vai trò quan trọng:

* Tạo đặc trưng mới: Một notebook trên Kaggle đã tạo đặc trưng TotalSF (tổng diện tích) bằng cách cộng GrLivArea và TotalBsmtSF, giúp cải thiện độ chính xác của mô hình.
* Mã hóa biến phân loại: Các biến như Neighborhood, BldgType, và KitchenQual thường được mã hóa bằng kỹ thuật one-hot encoding hoặc target encoding để phản ánh sự khác biệt về giá giữa các danh mục.
* Xử lý giá trị thiếu: Các cột như PoolQC, Alley, và Fence có tỷ lệ giá trị thiếu cao (trên 80%). Người dùng trên Kaggle thường điền giá trị "None" cho các cột này (vì giá trị thiếu thường có ý nghĩa, ví dụ: không có hồ bơi), hoặc loại bỏ các cột không cần thiết.
* Tuy nhiên, một số người dùng trên Kaggle nhận xét rằng Ames Housing có nhiều giá trị bị thiếu hơn Boston Housing, đòi hỏi kỹ thuật xử lý dữ liệu phức tạp hơn.

1. **Ứng dụng thực tế**

* Ngoài việc học tập và nghiên cứu, các kỹ thuật từ cuộc thi Kaggle đã được áp dụng trong ngành bất động sản. Ví dụ, các mô hình dự đoán giá nhà dựa trên Ames Housing có thể được sử dụng để định giá tự động (automated valuation models - AVMs) hoặc hỗ trợ các công ty bất động sản trong việc định giá tài sản.

1. **Bộ dữ liệu tương tự Ames Housing:**
   * Bộ dữ liệu Boston Housing (Harrison & Rubinfeld, 1978) là một tiền thân nổi tiếng, nhưng chỉ có 13 biến và 506 mẫu, ít chi tiết hơn so với Ames Housing (79 biến, 2.930 mẫu).
   * Bộ dữ liệu California Housing cũng được sử dụng rộng rãi, nhưng tập trung vào dữ liệu tổng hợp cấp khu vực hơn là chi tiết từng ngôi nhà như Ames Housing.
2. **Một số nghiên cứu tiêu biểu có thể kể đến:**

* *"Data Analysis on the Ames Housing Dataset"* (NYC Data Science Academy): Trình bày các đặc điểm của dữ liệu và cách tiếp cận trong xây dựng mô hình dự đoán giá nhà.
* *"Dự báo giá nhà bằng hồi quy Lasso và Ridge"* (RPubs): Ứng dụng các kỹ thuật hồi quy để tối ưu hóa việc dự đoán giá nhà.
* *"Predicting Housing Prices Using Machine Learning"* (Kaggle): Nghiên cứu sử dụng nhiều mô hình học máy khác nhau để tối ưu dự đoán giá nhà dựa trên tập dữ liệu Ames.

Phần 2: Review of Ames Housing (Đánh giá bộ dữ liệu Ames Housing)

Nội dung:

1. **Tổng quan về dữ liệu:**
   * Ames Housing chứa thông tin về 2.930 giao dịch nhà ở tại Ames, Iowa từ 2006-2010, với 79 biến độc lập và 1 biến phụ thuộc (SalePrice).
   * Các biến bao gồm cả định lượng (quantitative) như diện tích (GrLivArea), số phòng ngủ (BedroomAbvGr), và định tính (qualitative) như chất lượng tổng thể (OverallQual), tình trạng nhà (SaleCondition).
2. **Ưu điểm:**
   * Độ chi tiết: Với 79 biến, bộ dữ liệu cho phép phân tích sâu về nhiều khía cạnh của ngôi nhà, từ cấu trúc vật lý (diện tích, số tầng) đến yếu tố thẩm mỹ (chất lượng hoàn thiện).
   * Tính thực tế: Dữ liệu dựa trên giao dịch thực tế, phản ánh thị trường bất động sản tại một khu vực cụ thể.
   * Ứng dụng học máy: Phù hợp với các bài toán hồi quy, phân tích tương quan, và thử nghiệm các mô hình từ đơn giản đến phức tạp.
3. **Nhược điểm:**
   * Dữ liệu thiếu: Một số biến như PoolQC (chất lượng hồ bơi), MiscFeature (tính năng khác) có tỷ lệ thiếu dữ liệu cao (trên 90%), đòi hỏi kỹ thuật xử lý dữ liệu phức tạp.
   * Phạm vi giới hạn: Chỉ đại diện cho Ames, Iowa trong giai đoạn 2006-2010, không phản ánh xu hướng thị trường hiện tại (2025) hoặc các khu vực khác.
   * Phân phối lệch: Biến mục tiêu SalePrice có phân phối lệch phải, cần biến đổi để phù hợp với một số mô hình.
4. **Khả năng ứng dụng:**
   * Dự đoán giá nhà bằng các mô hình như XGBoost, Random Forest.
   * Phân tích các yếu tố quan trọng ảnh hưởng đến giá trị bất động sản (ví dụ: diện tích, chất lượng).
   * Dùng làm tài liệu giảng dạy trong các khóa học về phân tích dữ liệu và học máy.