TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

Tên bài tập:

**KHẢO SÁT BÀI BÁO CHO DỮ LIỆU AMES HOUSING**

Môn học: **Phương pháp nghiên cứu khoa học trong công nghệ thông tin**

Giảng viên hướng dẫn: **Đỗ Như Tài**

Nhóm lớp học: **Nhóm 2 (Thứ 7 tiết 4 + 5 )**

Sinh viên thực hiện: **Lê Khánh Hoàng** MSSV: **3122410125**

**Nguyễn Minh Thuận** MSSV: **3123410365**

**Nguyễn Tiến Trung** MSSV: **3123410396**

**Phan Hoàng Vũ** MSSV: **3123410426**

**TP.HCM, ngày 25 tháng 2 năm 2025**

**MỤC LỤC**

**Giải bài tập………………………………………………………………………3**

**RESEARCH PROPOSAL**

1. **TITLE**

Dự đoán Giá nhà Ames bằng các Kỹ thuật Hồi quy Tiên tiến

1. **INTRODUCTION (GIỚI THIỆU)**
   1. **Bối cảnh nghiên cứu**

Giá nhà liên tục biến động và có ảnh hưởng lớn đến nền kinh tế cũng như quyết định tài chính của người mua và nhà đầu tư. Việc dự đoán chính xác giá nhà có thể giúp:

* Người mua đưa ra quyết định hợp lý khi mua bất động sản.
* Nhà đầu tư có cơ sở để đánh giá lợi nhuận và rủi ro.
* Các nhà hoạch định chính sách có chiến lược phát triển nhà ở hợp lý hơn.

Ames, lowa, là một thành phố điển hình với dữ liệu bất động sản đầy đủ, phù hợp để nghiên cứu các yếu tố ảnh hưởng đến giá nhà và áp dụng các mô hình dự đoán tiên tiến**[1]**.

* 1. **Câu hỏi nghiên cứu**

1. Những yếu tố nào có ảnh hưởng mạnh nhất đến giá nhà tại Ames?
2. Mô hình hồi quy nào có thể dự đoán giá nhà chính xác nhất?
3. Ứng dụng các kỹ thuật học máy tiên tiến có cải thiện độ chính xác dự đoán không?
   1. **Mục tiêu nghiên cứu**

* Xác định các đặc trưng quan trọng nhất ảnh hưởng đến giá nhà.
* Ứng dụng các kỹ thuật học máy tiên tiến để xây dựng mô hình dự đoán.
* Đánh giá hiệu suất của các mô hình hồi quy khác nhau và chọn mô hình tối ưu nhất.

1. **LITERATURE REVIEW (TỔNG QUAN TÀI LIỆU)**
   1. **Các nghiên cứu trước đây**

* Rahidi et al. (2015) đã phân loại các yếu tố ảnh hưởng đến giá nhà thành đặc điểm vật lý, thiết kế kiến trúc, và vị trí địa lý**[2]**.
* Các nghiên cứu trước đây chủ yếu sử dụng hồi quy tuyến tính để dự đoán giá nhà, nhưng các mô hình phi tuyến tính như Gradient Boosting, XGBoost, và LightGBM chưa được khai thác đầy đủ**[3]**.
  1. **Khoảng trống nghiên cứu**
* Các phương pháp trước đây chưa tận dụng đầy đủ kỹ thuật feature engineering và regularization models để giảm thiểu sai số**[4]**.
* Chưa có nghiên cứu so sánh chi tiết giữa các mô hình học máy khác nhau trên tập dữ liệu Ames Housing**[5]**.

1. **RESEARCH METHODOLOGY (PHƯƠNG PHÁP NGHIÊN CỨU)**
   1. **DATASET (DỮ LIỆU NGHIÊN CỨU)**

* Nguồn dữ liệu: Dữ liệu từ Kaggle “House Prices: Advanced Regression Techniques” với 1460 căn nhà trong tập huấn luyện và 1459 căn nhà trong tập kiểm tra**[6]**.
* Các đặc trưng chính:
  + Numerical features: Diện tích tầng trệt, diện tích tầng hầm, số phòng ngủ, năm xây dựng.
  + Categorical features: Khu vực lân cận, chất lượng tổng thể, vật liệu xây dựng.
  1. **DATA PREPROCESSING (TIỀN XỬ LÝ DỮ LIỆU)**
* Xử lý giá trị thiếu:
  + Điền giá trị trung vị với các đặc trưng số.
  + Gán "None" cho các giá trị thiếu trong đặc trưng phân loại.
* Phát hiện và loại bỏ ngoại lệ ảnh hưởng đến mô hình.
* Chuyển đổi dữ liệu lệch bằng log transformation**[7]**.
* Chuẩn hóa dữ liệu để đảm bảo mô hình hoạt động ổn định.
  1. **MODEL SELECTION (LỰA CHỌN MÔ HÌNH)**
* Các mô hình hồi quy nâng cao được sử dụng để dự đoán giá nhà:

1. LASSO Regression: Để lựa chọn đặc trưng quan trọng và tránh overfitting.
2. Elastic Net Regression: Kết hợp ưu điểm của LASSO và Ridge Regression.
3. Gradient Boosting Regression: Mô hình học sâu có khả năng học phi tuyến tính.
4. XGBoost & LightGBM: Hai mô hình boosting mạnh mẽ, tối ưu hóa tốc độ và hiệu suất.
5. Stacked Model: Kết hợp nhiều mô hình để nâng cao độ chính xác.
   1. **MODEL EVALUATION (ĐÁNH GIÁ MÔ HÌNH)**

* Dùng 5-fold cross-validation để đánh giá hiệu suất mô hình**[8]**.
* Sử dụng Root-Mean-Square Logarithmic Error (RMSLE) để so sánh mô hình.
* Kiểm tra khả năng tổng quát hóa của mô hình trên tập kiểm tra.

1. **EXPECTED OUTCOMES (KẾT QUẢ MONG ĐỢI)**

* Xây dựng mô hình có thể dự đoán chính xác giá nhà với RMSLE thấp nhất.
* Xác định các yếu tố quan trọng nhất ảnh hưởng đến giá nhà tại Ames.
* Đưa ra đề xuất về chiến lược đầu tư bất động sản dựa trên kết quả dự đoán.

1. **TIMELINE (KẾ HOẠCH THỜI GIAN)**

|  |  |  |
| --- | --- | --- |
| Giai đoạn | Công việc | Thời gian (tuần) |
| 1 | Thu thập và tiền xử lý dữ liệu | 3 |
| 2 | Xây dựng mô hình hồi quy | 4 |
| 3 | Đánh giá mô hình bằng Cross-validation | 3 |
| 4 | Phân tích kết quả và viết báo cáo | 2 |

1. **REFERENCES (TÀI LIỆU THAM KHẢO)**

[1] Han, Y. (2023). Price Prediction of Ames Housing Through Advanced Regression Techniques.

[2] Rahadi, R. A. et al. (2015). Factors influencing the price of housing in Indonesia.

[3] Alfiyatin, A. N. et al. (2017). Modeling house price prediction using regression analysis.

[4] Zhu, D. X. & Wei, K. L. (2013). The Land Prices and Housing Prices.

[5] Jim, C. Y. & Chen, W. Y. (2009). Value of scenic views: Hedonic assessment of private housing.

[6] Marcelino, P. (2017). Comprehensive Data Exploration with Python.

[7] Serigne (2017). Stacked Regressions: Top 4% on LeaderBoard.

[8] Saha, S. (2022). XGBoost vs LightGBM: How Are They Different?.