ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN – TIN HỌC

----------o0o----------

ĐỒ ÁN MÔN HỌC:NHẬP MÔN KHOA HỌC DỮ LIỆU

GVHD: NGÔ MINH MẪN

TP. HỒ CHÍ MINH, THÁNG 06,NĂM 2024

ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN – TIN HỌC
BỘ MÔN : Nhập môn Khoa Học Dữ Liệu

----------o0o----------

1. GIỚI THIỆU

1.1 Tổng quan

Nhóm 23: 4 thành viên

- Nhóm trường: Trần Gia Huy
- Thành viên: Nguyễn Lê Đăng Khoa
- Thành viên: Đoàn Nhật Nam
- Thành viên: Phạm Bá Hoàng Anh

Project 1: House Price Prediction

1.2 Nhiệm vụ đề tài

Mô tả các nhiệm vụ của đồ án bao quét dữ liệu, tiền xử lý dữ liệu, kỹ thuật tính năng, mô hình hóa và đánh giá. Nhiệm vụ liên quan đến việc dự đoán giá nhà bằng mô hình học máy.

1.3 Phân chia công việc trong nhóm

Các thành viên trong nhóm thực hiện các công việc sau:

- Trần Gia Huy: xử lý Large Language Model để lấy thông tin trong mô tả, data preprocessing, feature engineeing, eda và MLP từ Pytorch
- Nguyễn Lê Đăng Khoa: crawl data, feature engineering, methodology
- Đoàn Nhật Nam : crawl data, eda , data preprocessing, methodology
- Phạm Bá Hoàng Anh: methodology

2. Phương pháp:

- Data collecting: nhóm chúng em đề xuất cào dữ liệu từ web https://batdongsan.vn/ cụ ở trên địa bàn thành phố Hồ Chí Minh, chiết xuất các features như giá nhà , số phòng ngủ, số WC, địa chỉ và mô tả
- Sau đấy, nhóm chúng em sẽ sử dụng Large Language Model để chiết xuất các feature như giá nhà, số phòng ngủ, số WC, địa chỉ, bề rộng, bề dài, có mặt tiền hay là không và số tầng .Đối với các feature có sẵn từ việc crawl data từ web thì nhóm chúng em sẽ sử dụng dữ liệu có được từ LLM để fill missing value, còn đối với những features chưa có được từ crawl data mà chỉ có được bằng cách sử dụng LLM trên mô tả thì sẽ được sử dụng cho việc huấn luyện model như bề rộng, bề dài,...

- Data preprocessing: Nhóm chúng em sẽ xử lý các dữ liệu sai lệch trong Price như 3900000000 tỷ, 500000000 tỷ, 500 triệu tỷ,... bằng cách lấy đơn vị ra và xử lý như trong code. Loại bỏ các ký tự thừa trong các features 'Area','Bedrooms','WC'.Chuyển đổi dữ liệu sang dạng số và xử lý missing value với mỗi feature là khác nhau : như với số phòng ngủ,wc thì sẽ sử dụng KNN có trọng số, đối với số tầng thì sẽ là mặc định là 1 đối với missing values,...
- Feature engineering: bên cạnh việc xử lý missing value, thì nhóm chúng em cũng One-Hot encoding đối với một số features như 'Khu_Vuc' và 'Facade'. Thêm feature như tổng diện tích, xử lý outliner,loại bỏ feature có correalation và mi scores thấp đối với target value để cải thiện hiệu suất model,...
- Training: nhóm chúng em sử dụng các model có performance tốt đối với data như Linear Regression, DecisionTree, Random Forest và có thử sư

3. Đánh giá:

- Trong đồ án nhóm em có sử dụng 2 mô hình mà nhóm em cảm thấy phù hợp nhất với tập dữ liệu mà nhóm thu thập được là Linear Regression và GradientBoostingRegressor:

- **Kết quả so sánh**

a. Độ chính xác:

- **Linear Regression:**
  - Ưu điểm: Đơn giản, dễ hiểu, và dễ triển khai. Hiệu suất tốt với dữ liệu tuyến tính hoặc gần tuyến tính.
  - Nhược điểm: Hiệu suất kém với dữ liệu phi tuyến tính hoặc có sự phức tạp cao. Độ chính xác thấp hơn so với GradientBoostingRegressor trong nhiều trường hợp.
- **GradientBoostingRegressor:**
  - Ưu điểm: Độ chính xác cao hơn, đặc biệt với dữ liệu phi tuyến tính. Khả năng xử lý tốt các mối quan hệ phức tạp trong dữ liệu.
  - Nhược điểm: Có thể dễ bị overfitting nếu không điều chỉnh đúng các siêu tham số.

b. Hiệu suất dự đoán:

- **Linear Regression:**
  - Phân bố dữ liệu: Các điểm dữ liệu phân bố xung quanh đường y = x nhưng sự phân tán rộng hơn, đặc biệt ở các giá trị cực trị.
  - Độ chính xác: Hiệu suất không tốt bằng dữ liệu huấn luyện, nhiều điểm dự đoán nằm xa giá trị thực tế.
- **GradientBoostingRegressor:**
  - Phân bố dữ liệu: Các điểm dữ liệu cũng phân bố xung quanh đường y = x, nhưng ít phân tán hơn so với Linear Regression trên dữ liệu kiểm tra.
  - Độ chính xác: Hiệu suất tốt hơn trên dữ liệu kiểm tra, nhiều điểm nằm gần giá trị thực tế hơn.

4. Kết luận

- Yêu cầu đặt ra cho bài làm
    - Hoàn thành được các yêu cầu đặt ra của đề bài
    - Kết hợp được crawl data và llm trong việc collect data
- **Phân tích**
    - Có giải thích code chi tiết của từng phần trong Jupyter Notebook