BTLT - NM TTNT - TUẦN 3

21280099 - Nguyễn Công Hoài Nam

Ngày 23 tháng 12 năm 2023

I. Giới thiệu

1. Đề bài

Em hãy chọn 1 project về phân tích dữ liệu mà em thích, nêu ra kế hoạch thực hiện (Planning flow) từ bước thu thập dữ liệu và sau đó phát triển tính năng và nghiệm thu/pilot.

2. Project được chọn

Phân tích dữ liệu nhà ở để dự đoán dự đoán giá nhà tại Việt Nam

II. Các bước thực hiện

1. Problem Definition - Xác định vấn đề

- **Mục tiêu chính:** Dự đoán dự đoán giá nhà tại Việt Nam dựa trên các đặc trưng như địa chỉ, diện tích, số phòng ngủ, nhà vệ sinh, số tầng, tiện nghi,...
- **Yêu cầu:** Một bộ dữ liệu đủ lớn và đa dạng, chứa các thông tin cần thiết

2. Data Gathering and Preparation - Chuẩn bị và thu thập dữ liệu

a. Data Access - Truy cập dữ liệu

- **Nguồn dữ liệu:** Dữ liệu được lấy từ website: *batdongsan.vn*
- **Quy trình thu thập dữ liệu:** Xây dựng thuật toán cào web sử dụng thư viện BeautifulSoup và Python.

b. Data Sampling - Lấy mẫu dữ liệu

- **Xác định kích thước mẫu:** Quyết định số lượng mẫu cần thiết để đại diện cho thị trường bất động sản tại Việt Nam (lấy cỡ mẫu = 10000)
- **Phương pháp lấy mẫu:** Lấy các nhà ở Việt Nam, lựa chọn những trường đặc trưng hữu ích như diện tích, số phòng,...

c. Data Transformation

- **Làm sạch và tiền xử lý dữ liệu:** Khử giá trị ngoại lai (outliers), giá trị null, missing value,..
- **Tìm thêm đặc trưng mới**: Dựa trên dữ liệu có sẵn, tạo ra các đặc trưng mới dựa vào toạ độ tìm các tiện ích xung quang và khoảng cách của chúng,...

Ngoài ra có thể EDA (Exploratory Data Analysis) để có cái nhìn tổng thể về dữ liệu, vẽ biểu đồ thể hiện tương quan của các đặc trưng, ...

3. Model Building and Evaluation

a. Create Model

- Chọn mô hình phù hợp: Dựa trên bản chất của bài toán (dự đoán giá nhà), có thể sử dụng các mô hình như Linear Regression, Random Forest, Gradient Boosting, hoặc Neural Networks.

- Xây dựng mô hình: Sử dụng dữ liệu đã được chuẩn bị để huấn luyện mô hình dự đoán giá nhà.

b. Test Model

- **Tách dữ liệu thành tập huấn luyện và tập kiểm tra:** Phân chia dữ liệu thành tập huấn luyện và tập kiểm tra để đánh giá hiệu suất của mô hình.
- **Kiểm thử mô hình:** Sử dụng tập dữ liệu kiểm tra để đánh giá khả năng dự đoán của mô hình trên dữ liệu mới.

c. Evaluate and Interpret Model

- **Đánh giá hiệu suất:** Sử dụng các độ đo như RMSE (Root Mean Squared Error), R-squared, hay MAE (Mean Absolute Error) để đánh giá độ chính xác của mô hình.
- **Phân tích và giải thích mô hình:** Xác định các yếu tố quan trọng ảnh hưởng đến việc dự đoán giá nhà, hiểu rõ hơn về cách mô hình quyết định dự đoán và tác động của từng đặc trưng.

4. Knowledge Deployment

a. Model Apply

- **Áp dụng mô hình:** Sử dụng mô hình đã xây dựng để dự đoán giá nhà trên dữ liệu thực tế hoặc mới.

b. Custom Reports

- **Tạo báo cáo tùy chỉnh:** Tạo ra các báo cáo hoặc biểu đồ tùy chỉnh để trình bày kết quả dự đoán và phân tích từ mô hình.

c. External Applications

- **Ứng dụng bên ngoài:** Kết nối mô hình dự đoán giá nhà với các ứng dụng hoặc hệ thống bên ngoài để sử dụng thông tin dự đoán trong môi trường khác nhau (chẳng hạn làm website dự đoán giá nhà)

III. Kết luận

Tóm lại chúng ta đã thực hiện một chuỗi các bước từ việc thu thập dữ liệu đến xây dựng mô hình, đánh giá và triển khai tri thức để dự đoán giá nhà tại Việt Nam. Quá trình này giúp chúng ta hiểu rõ hơn về yếu tố ảnh hưởng đến giá nhà và cung cấp cơ sở cho việc đưa ra các quyết định thông minh trong thị trường bất động sản.