Đề thi:

DATA PRE-PROCESSING AND DATA ANALYSIS

Ngày thi : 27/03/2024

**** Học viên tạo 1 thư mục là **DL04_HoVaTen**, lưu tất cả bài làm vào để nộp chấm điểm ****

**** Học viên được sử dụng tài liệu ****

**** Với mỗi câu, sử dụng Markdown để mô tả yêu cầu ****

Phần 1 : Đọc tập tin dữ liệu dự báo giá nhà (1đ)

1. Đọc tập tin dữ liệu housing-prices-dataset.csv
2. Xem thông tin sơ bộ : shape/head/tail/info
3. Kiểm tra dữ liệu bị trùng và xử lý

Phần 2 : Phân tích EDA (4đ)

1. Chọn các biến sau đây để phân tích : 'LotShape', 'Street', 'HouseStyle', 'LotArea', 'YearBuilt', '1stFlrSF', '2ndFlrSF', 'FullBath', 'BedroomAbvGr', 'TotRmsAbvGrd', 'SalePrice'

Biến phụ thuộc là biến SalePrice

2. Xác định chính xác các biến số, các biến phân loại (trong các biến ở câu 1)
3. Kiểm tra dữ liệu bị thiếu và xử lý
4. Phân tích 1 biến (cho nhận xét)
5. Phân tích 2 biến (cho nhận xét)
6. Kiểm tra và xóa các outlier

Phần 3 : Feature Engineering (1.5đ)

1. Chuẩn hóa các biến phân loại input bằng one-hot encoder/ label encoder
2. Chuẩn hóa các biến số input bằng Log Normalization/ Standard Scaler/ min-max Scaler/ Robust Scaler

Phần 4 : Tạo mô hình Linear Regression và đánh giá (2đ)

1. Chia tập dữ liệu thành 2 tập train và test (test size : 0.2)
2. Tạo mô hình Linear Regression và huấn luyện với tập train
3. Đánh giá mô hình (score : r squared, mse, mae ; có vẽ biểu đồ) trong cả 3 trường hợp : Full, Train, Test. Cho nhận xét

Phần 5 : Cải tiến hiệu suất mô hình (1.5đ)

Học viên có thể chọn các biến tùy ý để áp dụng cho mô hình

1. Giải pháp 1: tạo hàm đa thức bậc 2 (từ các biến trong câu 1 phần 2), cho mô hình học lại và đánh giá
2. Giải pháp 2: sử dụng SelectKBest (sklearn) để chọn các feature có score cao nhất, cho mô hình học lại và đánh giá
3. Giải pháp 3: loại bỏ đa cộng tuyến trong các biến input , cho mô hình học lại và đánh giá