BỘ NÔNG NGHIỆP VÀ PHÁT TRIỂN NÔNG

THÔN PHÂN HIỆU TRƯỜNG ĐẠI HỌC THỦY

LỢI

BỘ MÔN CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỌC MÁY

ĐỀ TÀI: 3A Superstore (Market Orders Data-CRM)

DỰ ĐOÁN PROFIT CÁC ĐƠN HÀNG

GV hướng dẫn: ThS.Vũ Thị Hạnh

Lớp: S25-64CNTT

Sinh viên thực hiện	Mã số sinh viên
Lâm Chí Trung	2251068272

TP.HCM, tháng 10 năm 2025

1. Giới thiệu đề tài

Đề tài 3A Superstore (Market Orders Data – CRM) được thực hiện nhằm phân tích dữ liệu bán hàng của một chuỗi siêu thị, từ đó ứng dụng các kỹ thuật khai phá dữ liệu và học máy để dự đoán lợi nhuận (Profit) cho từng đơn hàng. Hệ thống CRM (Customer Relationship Management – Quản lý quan hệ khách hàng) là nền tảng giúp doanh nghiệp quản lý thông tin khách hàng, theo dõi hoạt động bán hàng và hỗ trợ ra quyết định. Việc phân tích dữ liệu CRM có ý nghĩa quan trọng trong việc tối ưu chiến lược kinh doanh.

Ngoài việc phân tích dữ liệu CRM, đề tài còn xây dựng **ứng dụng web (Dashboard)** giúp người dùng trực quan hóa dữ liệu và dự đoán lợi nhuận trực tiếp bằng mô hình học máy.

2. Giới thiệu dữ liệu (Dataset)

Tập dữ liệu được sử dụng trong đề tài có tên **Superstore.xls**, đây là bộ dữ liệu mẫu phổ biến trong lĩnh vực *Business Intelligence* và *CRM (Customer Relationship Management)*. Dữ liệu ghi nhận các **đơn hàng bán lẻ của một chuỗi siêu thị** trong nhiều khu vực khác nhau, bao gồm thông tin về sản phẩm, khách hàng, hình thức vận chuyển và doanh thu.

Cụ thể, tập dữ liệu có khoảng **65.000 dòng** (mỗi dòng là một đơn hàng) và hơn **21 cột** mô tả chi tiết các khía cạnh sau:

- **Thông tin đơn hàng:** Order ID, Order Date, Ship Date, Ship Mode
- **Thông tin khách hàng:** Customer ID, Customer Name, Segment
- **Khu vực và địa lý:** Country, City, State, Region
- **Thông tin sản phẩm:** Category, Sub-Category, Product Name
- **Chỉ số kinh doanh:** Sales, Quantity, Discount, Profit

Trong đó, biến mục tiêu (**target variable**) của bài toán là **Profit** – biểu thị **lợi nhuận thu được từ mỗi đơn hàng**.
Các cột còn lại được xem là **đặc trưng đầu vào (features)** phục vụ cho việc huấn luyện mô hình học máy.

Dữ liệu này phản ánh tương đối đầy đủ quy trình quản lý đơn hàng trong hệ thống CRM, giúp phân tích và dự đoán hiệu quả hoạt động kinh doanh của doanh nghiệp.

Dữ liệu được chia làm 80% tập huấn luyện và 20% tập kiểm tra.
Một số cột không cần thiết (như ID, ngày tháng) đã được loại bỏ trong giai đoạn tiền xử lý.

3. Mục tiêu và dữ liệu sử dụng

Mục tiêu của đề tài là xây dựng mô hình học máy có khả năng dự đoán lợi nhuận của từng đơn hàng dựa trên các đặc trưng như: doanh số (Sales), số lượng (Quantity), mức giảm giá (Discount), phương thức giao hàng (Ship Mode), phân khúc khách hàng (Segment), khu vực (Region), loại sản phẩm (Category) và nhóm sản phẩm (Sub-Category). Dữ liệu được sử dụng là tập Superstore.xls, chứa thông tin chi tiết về các đơn hàng trong hệ thống CRM của doanh nghiệp.

4. Quy trình thực hiện

Quy trình thực hiện gồm các bước:

Tiền xử lý dữ liệu:

- Xử lý giá trị thiếu, chuẩn hóa dữ liệu số (`StandardScaler`), mã hóa biến phân loại (`OneHotEncoder`) bằng `ColumnTransformer`.
- Xây dựng `Pipeline` tự động kết hợp bước tiền xử lý và huấn luyện mô hình.

Xây dựng và huấn luyện mô hình:

- Linear Regression
- Decision Tree Regressor (max_depth=8)
- Random Forest Regressor (n_estimators=100, max_depth=10)
- XGBoost Regressor (n_estimators=100, learning_rate=0.1)

Đánh giá mô hình:

- Sử dụng các chỉ số R^2, MAE, RMSE để so sánh.
- Ghi nhận mô hình tốt nhất là **XGBoost ($R^2 \approx 0.983$)**.

Triển khai Dashboard Web:

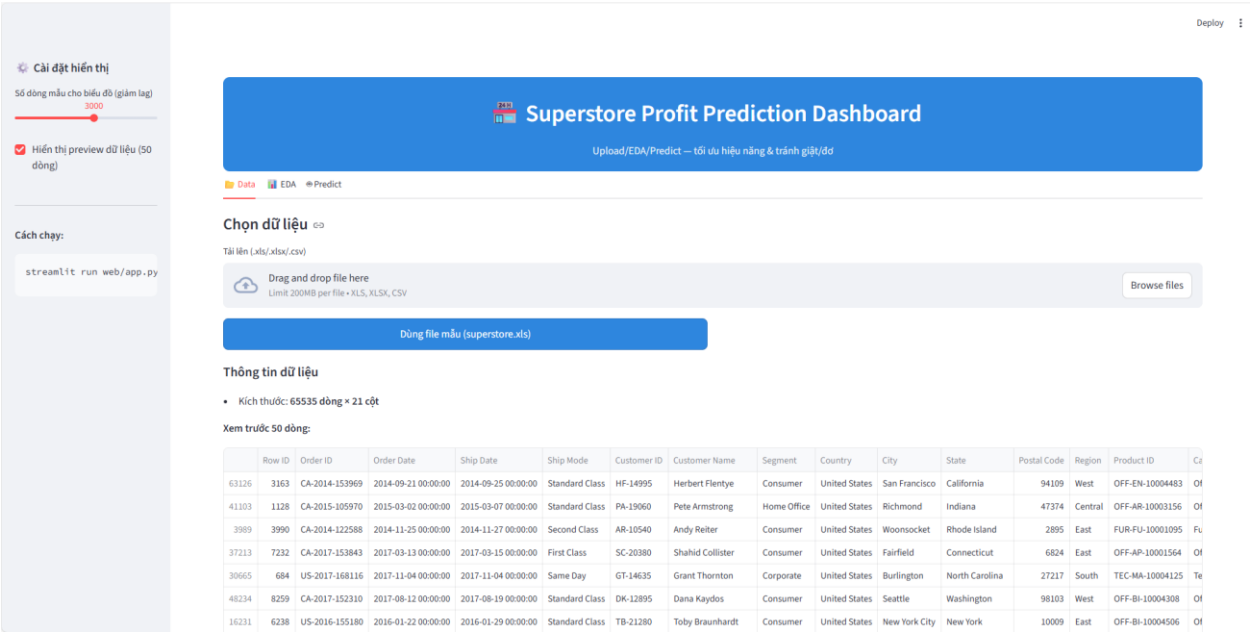
- Sử dụng **Streamlit** để hiển thị các biểu đồ EDA và giao diện nhập liệu dự đoán lợi nhuận.
- Giao diện gồm 3 tab: **Data – EDA – Predict**.
- Người dùng có thể chọn mô hình để dự đoán (Linear, Tree, Forest, XGBoost).

5. Kết quả và đánh giá

Mô hình	R2	MAE	RMSE
Linear Regression	0.885	32.31	80.19
Decision Tree	0.953	22.47	51.46
Random Forest	0.983	14.30	31.12
XGBoost	0.983	15.45	30.82

Giao diện web:
Ứng dụng web có 3 tab chức năng:

- **Data:** cho phép người dùng tải file dữ liệu hoặc dùng mẫu.
- **EDA:** hiển thị biểu đồ doanh thu, lợi nhuận, mối quan hệ Sales–Profit.
- **Predict:** nhập các thông số đầu vào và chọn mô hình để dự đoán Profit.



6. Kết luận và hướng phát triển

Đề tài đã hoàn thiện pipeline tiền xử lý, huấn luyện, đánh giá, và triển khai thành **ứng dụng web hoàn chỉnh**.
Việc áp dụng mô hình XGBoost và RandomForest giúp cải thiện độ chính xác rõ rệt so với Linear và DecisionTree.
Dashboard web giúp người dùng **tương tác, so sánh mô hình và dự đoán nhanh**, đáp ứng đúng tiêu chí điểm cộng của môn học.

Hướng phát triển:

- Tích hợp chức năng lưu lịch sử dự đoán.

- Đưa ứng dụng lên Streamlit Cloud hoặc HuggingFace Spaces.
- Mở rộng dự đoán doanh thu hoặc khuyến mãi tối ưu.