Bài thực hành số 8. Sử dụng thư viện ML

Mục tiêu: sử dụng được các lớp cơ bản của ML: Transform, Estmator, Pipeline để xây dựng mô hình học máy và đánh giá mô hình.

Nội dung thực hành:

Dự đoán khách hàng rời ngân hàng

Cho bộ dữ liệu "Churn_Modelling.csv" mô tả các thông tin khách hàng của một ngân hàng, bao gồm:

RowNumber: số thứ tự của mỗi dòng trong bộ dữ liệu

CustomerId: mã số khách hàng

Surname: họ của khách hàng

CreditScore: điểm tín dụng của khách hàng

Geography: quốc gia của khách hàng

Gender: giới tính của khách hàng

Age: tuổi của khách hàng

Tenure: số năm khách hàng đã sử dụng dịch vụ của ngân hàng

Balance: số tiền trong tài khoản của khách hàng

NumOfProducts: số lượng sản phẩm mà khách hàng đã mua từ ngân hàng

HasCrCard: có sở hữu thẻ tín dụng không (1 nếu có, 0 nếu không)

IsActiveMember: khách hàng có đang hoạt động trong ngân hàng không (1 nếu đang hoạt động, 0 nếu không)

EstimatedSalary: ước tính thu nhập của khách hàng

Exited: khách hàng đã rời đi hay chưa (1 nếu đã rời đi, 0 nếu chưa)

Yêu cầu:

a) Sử dụng PySpark để đọc dữ liệu vào DataFrame.
b) Thực hiện một số thống kê, trực quan hóa để hiểu dữ liệu.
c) Tiền xử lý dữ liệu, bao gồm: loại bỏ cột RowNumber, CustomerID, chuyển đổi giá trị chuỗi thành số, chuyển đổi các biến độc lập thành vector.
d) Chia dữ liệu thành tập huấn luyện và tập kiểm tra với tỉ lệ 70/30.
e) Sử dụng Logistic Regression để huấn luyện mô hình dự đoán khách hàng có rời khỏi ngân hàng không?
f) Đánh giá hiệu suất của mô hình trên tập kiểm tra bằng độ chính xác, độ phủ, độ chính xác cân bằng, F1 score và AUC.
g) Tạo pipeline để xây dựng mô hình từ bước chuẩn hóa dữ liệu đến chọn mô hình học máy. Sử dụng pipeline để huấn luyện mô hình từ dữ liệu huấn luyện.
h) Lưu mô hình xây dựng bao gồm dữ liệu đã huấn luyện.
i) Mở lại mô hình đã lưu và dự đoán lại cho dữ liệu test.

Lựa chọn thuộc tính xây dựng mô hình

Từ việc hiểu dữ liệu, phân tích độ tương quan giữa các thuộc tính hãy chọn các thuộc tính cần thiết cho việc xây dựng mô hình.

Xây dựng mô hình với các thuộc tính đã chọn.

Đánh giá mô hình và so sánh kết quả đánh giá với mô hình xây dựng ở trên.

Thay đổi tham số của mô hình

Thay đổi tham số của mô hình LogisticRegression như: maxIter, regParam rồi so sánh với kết quả ban đầu.
