[DA_Project TIMA] Task 12_Xây dựng mô hình phân loại

Mô hình phân loại (Classification Model) là một loại mô hình học máy (machine learning model) được sử dụng để phân loại dữ liệu vào các nhóm hoặc lớp (class) khác nhau dựa trên các đặc trưng (features) của dữ liệu đầu vào. Mục tiêu của mô hình phân loại là dự đoán nhãn (label) của một đối tượng (ví dụ: khách hàng, sản phẩm, sự kiện, v.v.) dựa trên các đặc trưng có sẵn.

Quy trình xây dựng mô hình phân loại

Bước 1: Xác định bài toán phân loại và biến mục tiêu.

Bước 2: Thu thập và chuẩn bị dữ liệu.

Bước 3: Chia dữ liệu thành tập huấn luyện và kiểm tra.

Bước 4: Lựa chọn mô hình phân loại.

Bước 5: Huấn luyện mô hình.

Bước 6: Đánh giá mô hình và các chỉ số hiệu suất.

Bước 7: Cải thiện mô hình.

Bước 8: Triển khai mô hình vào sản xuất và theo dõi.

Quy trình chi tiết

1.1 Thu thập và xử lý dữ liệu

Dữ liệu phải được chuẩn bị kỹ lưỡng, bao gồm việc làm sạch, xử lý các giá trị thiếu, chuẩn hóa các đặc trưng và mã hóa các giá trị phân loại nếu cần thiết.

1.2 Chia dữ liệu

Dữ liệu được chia thành hai phần: tập huấn luyện (training set) và tập kiểm tra (test set).

1.3 Chọn mô hình

Lựa chọn mô hình phù hợp với yêu cầu bài toán

1.4 Huấn luyện mô hình

Sử dụng tập huấn luyện để xây dựng mô hình, tìm ra các tham số tối ưu cho mô hình.

1.5 Đánh giá mô hình

- Sử dụng tập kiểm tra để đánh giá hiệu suất của mô hình. Các chỉ số thường dùng:
  - Accuracy: Tỷ lệ phân loại chính xác.
  - Precision: Độ chính xác khi dự đoán lớp tích cực.
  - Recall: Tỷ lệ tìm được các mẫu thuộc lớp tích cực.
  - F1-score: Trung bình hài hòa của precision và recall.

1.6 Triển khai mô hình

- Sau khi đánh giá, mô hình có thể được triển khai để dự đoán trên các dữ liệu thực tế.

Các loại mô hình phân loại phổ biến:

2. Logistic Regression

- Là một mô hình phân loại cơ bản, được sử dụng khi đầu ra là nhãn nhị phân (ví dụ: có nợ xấu hoặc không có nợ xấu).

- Cách thức hoạt động: Sử dụng một hàm logistic (hay sigmoid) để dự đoán xác suất thuộc về một lớp cụ thể (0 hoặc 1).

3. K-Nearest Neighbors (KNN)

- Đây là một phương pháp phân loại không tham số. Mô hình này phân loại một đối tượng dựa trên nhãn của các đối tượng k láng giềng gần nhất.
- Cách thức hoạt động: Tính toán khoảng cách (ví dụ: Euclidean distance) giữa điểm cần phân loại và các điểm khác trong dữ liệu. Đối tượng được gán nhãn của lớp mà có nhiều đối tượng k láng giềng nhất.

4. Decision Trees

- Là một mô hình phân loại sử dụng cây quyết định để phân loại các đối tượng. Mô hình này tạo ra các câu hỏi phân chia dữ liệu theo các đặc trưng để dự đoán nhãn.
- Cách thức hoạt động: Cây quyết định chia nhỏ không gian đặc trưng thành các vùng khác nhau, mỗi vùng ứng với một nhãn phân loại.

5. Random Forest

- Là một mô hình phân loại dựa trên tập hợp nhiều cây quyết định. Nó xây dựng một tập hợp các cây quyết định (tập hợp này gọi là "rừng cây") và sử dụng kết quả của các cây này để đưa ra dự đoán.
- Cách thức hoạt động: Mỗi cây quyết định trong rừng sẽ đưa ra một dự đoán, và Random Forest sẽ chọn lớp có số lượng dự đoán lớn nhất từ tất cả các cây.

6. Support Vector Machines (SVM)

- Là một mô hình phân loại mạnh mẽ có thể xử lý dữ liệu không tuyến tính.
- Cách thức hoạt động: Tìm một siêu phẳng tối ưu phân chia các lớp dữ liệu sao cho khoảng cách giữa siêu phẳng và các điểm dữ liệu gần nhất (gọi là margin) là lớn nhất.

7. Naive Bayes

- Là một mô hình phân loại dựa trên định lý Bayes và giả định rằng các đặc trưng là độc lập với nhau.
- Cách thức hoạt động: Dự đoán xác suất của các lớp nhãn dựa trên các đặc trưng đã cho, sau đó chọn lớp có xác suất cao nhất.

8. Neural Networks (Mạng nơ-ron)

- Là mô hình phân loại mạnh mẽ, đặc biệt trong các bài toán phức tạp, với các lớp ẩn giúp xử lý dữ liệu phi tuyến tính.
- Cách thức hoạt động: Mạng nơ-ron gồm nhiều lớp (input, hidden, output) và sử dụng thuật toán học để điều chỉnh trọng số các kết nối giữa các lớp nhằm tối ưu hóa dự đoán.

Áp dụng kiến thức xây dựng các mô hình sau

1. Phân loại trạng thái tín dụng của khách hàng

- **Biến mục tiêu**: Trạng thái
- **Cách tính**:
  - Tính trạng thái tín dụng của khách hàng dựa vào các thông tin như số tiền còn lại, lịch sử thanh toán, và nợ xấu.
  - Các giá trị có thể là:
    - "Đang vay xong" (Khách hàng đã thanh toán toàn bộ nợ)
    - "Kết thúc" (Khoản vay đã hết hạn và thanh toán đầy đủ)
    - "Đang vay" (Khách hàng vẫn còn nợ)
    - "Nợ xấu" (Khách hàng vi phạm hợp đồng, không trả nợ đúng hạn)

2. Phân loại khách hàng có nợ xấu

- **Biến mục tiêu**: HasBadDebt
- **Cách tính**:
  - Nếu khách hàng có lịch sử nợ xấu (dựa vào cột "LongestOverdue" và "CreditInfo"), biến mục tiêu là **1** (Có nợ xấu).
  - Nếu không có lịch sử nợ xấu, giá trị là **0** (Không có nợ xấu).

3. Phân loại khách hàng có trả nợ trễ

- **Biến mục tiêu**: HasLatePayment
- **Cách tính**:
  - Nếu khách hàng có bất kỳ lần trễ hạn nào (dựa vào cột "LongestOverdue" hoặc các thông tin liên quan đến việc trễ hạn thanh toán), biến mục tiêu là **1** (Trễ hạn).
  - Nếu không có trễ hạn, giá trị là **0** (Không trễ hạn).

4. Phân loại khách hàng theo thu nhập

- **Biến mục tiêu**: Salary
- **Cách tính**:
  - Thu nhập được phân loại thành các nhóm:
    - Thu nhập thấp: Dưới 5 triệu.
    - Thu nhập trung bình: Từ 5 triệu đến 20 triệu.
    - Thu nhập cao: Trên 20 triệu.
  - Các giá trị phân loại này giúp phân nhóm khách hàng theo khả năng tài chính.

5. Dự đoán khả năng vay thêm của khách hàng

- **Biến mục tiêu**: NumberOfLoans
- **Cách tính**:
  - Tính toán số lượng khoản vay hiện tại của khách hàng. Nếu khách hàng có nhiều khoản vay (ví dụ: >2 khoản vay), có thể dự đoán rằng khách hàng có khả năng vay thêm trong tương lai.

6. Phân loại khách hàng theo loại sản phẩm tín dụng

- **Biến mục tiêu**: ProductCreditName
- **Cách tính**:
  - Phân loại khách hàng theo nhóm sản phẩm tín dụng mà họ sử dụng (Vay mua nhà, vay tiêu dùng, vay ô tô, vay kinh doanh).
  - Mỗi loại sản phẩm sẽ là một lớp phân loại riêng biệt.

7. Phân loại khách hàng theo tình trạng tín dụng (Thấp, Trung bình, Cao)

- **Biến mục tiêu**: TS_CREDIT_SCORE_V2
- **Cách tính**:
  - Dựa trên điểm tín dụng của khách hàng, phân thành các nhóm:
    - Điểm tín dụng thấp (dưới 500)
    - Điểm tín dụng trung bình (từ 500 đến 700)
    - Điểm tín dụng cao (trên 700)

8. Dự đoán khả năng khách hàng thanh toán đầy đủ nợ

- **Biến mục tiêu**: Trạng thái
- **Cách tính**:
  - Dự đoán trạng thái khoản vay của khách hàng: liệu họ có hoàn tất khoản vay đúng hạn hay không.
  - Dựa vào lịch sử thanh toán, số tiền nợ còn lại và các chỉ số tài chính.

9. Phân loại khách hàng theo độ tuổi

- **Biến mục tiêu**: Birthday
- **Cách tính**:
  - Tính độ tuổi của khách hàng từ ngày sinh, sau đó phân loại vào các nhóm độ tuổi:
    - 18-25 tuổi
    - 26-40 tuổi
    - 41-60 tuổi
    - Trên 60 tuổi

10. Dự đoán khả năng gia hạn khoản vay

- **Biến mục tiêu**: Trạng thái
- **Cách tính**:
  - Dựa trên thông tin khoản vay, lịch sử thanh toán và các yếu tố như mức thu nhập, phân tích khả năng khách hàng sẽ yêu cầu gia hạn khoản vay trong tương lai.

11. Phân loại khách hàng theo mức độ rủi ro tín dụng

- **Biến mục tiêu**: HasLatePayment, HasBadDebt
- **Cách tính**:
  - Dựa vào các yếu tố như số lần trễ hạn và tình trạng nợ xấu, khách hàng được phân loại thành các nhóm rủi ro tín dụng:
    - Rủi ro thấp
    - Rủi ro trung bình
    - Rủi ro cao

12. Phân loại khách hàng theo mức độ ổn định tài chính

- **Biến mục tiêu**: JobName, Salary
- **Cách tính**:
  - Dựa vào nghề nghiệp và mức thu nhập của khách hàng, phân loại thành các nhóm ổn định tài chính:
    - Nghề nghiệp ổn định (công chức, viên chức)
    - Nghề nghiệp tự do (doanh nhân, lao động tự do)

13. Phân loại khách hàng theo khu vực địa lý

- **Biến mục tiêu**: CityName, DistrictName
- **Cách tính**:
  - Phân loại khách hàng theo khu vực sống (thành phố, quận, phường).
  - Các khu vực có thể có mức độ rủi ro tín dụng khác nhau, giúp ngân hàng phân loại chiến lược tín dụng.

14. Phân loại khách hàng theo hình thức cư trú

- **Biến mục tiêu**: Hình thức cư trú
- **Cách tính**:
  - Phân loại khách hàng theo các nhóm như "Sở hữu nhà", "Thuê nhà", "Sống cùng gia đình".
  - Các nhóm này có thể giúp đánh giá khả năng tài chính và sự ổn định của khách hàng.

15. Phân loại khách hàng theo mức độ sử dụng tín dụng

- **Biến mục tiêu**: Số tiền đăng ký vay ban đầu, Tiền giải ngân
- **Cách tính**:
  - Phân loại khách hàng theo số tiền vay mà họ yêu cầu và số tiền thực tế đã được giải ngân. Các nhóm có thể là:

        - Mức vay thấp (< 50 triệu)
        - Mức vay trung bình (50 triệu - 200 triệu)
        - Mức vay cao (> 200 triệu)

16. Dự đoán khả năng khách hàng tiếp tục sử dụng sản phẩm tín dụng

- **Biến mục tiêu**: ProductCreditName
- **Cách tính**:
    - Dự đoán liệu khách hàng sẽ tiếp tục sử dụng sản phẩm tín dụng (vay mua nhà, vay tiêu dùng, v.v.) dựa trên lịch sử vay và khả năng trả nợ.

17. Phân loại khách hàng theo nhóm nghề nghiệp

- **Biến mục tiêu**: JobName
- **Cách tính**:
    - Phân loại khách hàng theo nghề nghiệp để xác định khả năng tài chính và tiềm năng vay vốn. Các nhóm có thể là:
        - Công chức, viên chức
        - Doanh nhân
        - Lao động tự do

18. Dự đoán khả năng vay thêm từ khách hàng hiện tại

- **Biến mục tiêu**: NumberOfLoans
- **Cách tính**:
    - Dựa trên số lượng khoản vay hiện tại của khách hàng, dự đoán liệu họ sẽ tiếp tục vay thêm trong tương lai (kết hợp với các yếu tố như thu nhập, lịch sử tín dụng).

19. Phân loại khách hàng theo nhóm gia đình

- **Biến mục tiêu**: RelativeFamilyName, FullNameFamily
- **Cách tính**:
    - Dựa vào thông tin về gia đình, phân loại khách hàng theo nhóm gia đình (tên người thân, mối quan hệ gia đình) để đánh giá sự ổn định tài chính và khả năng trả nợ.

20. Phân loại khách hàng theo sự thay đổi về lương

- **Biến mục tiêu**: Salary, ReceiveYourIncomeSalary
- **Cách tính**:
    - Dựa trên mức lương hiện tại và sự thay đổi lương (tăng, giảm), phân loại khách hàng thành các nhóm có sự thay đổi thu nhập:
        - Lương ổn định
        - Lương tăng
        - Lương giảm