Lab 3 - PHÁT HIỆN BẤT THƯỜNG TRONG DỮ LIỆU TÀI CHÍNH

1 Tập dữ liệu tài chính

1.1 Giới thiệu

- Bộ dữ liệu được cung cấp trong lab này là về báo cáo tài chính cho việc mua trả góp các sản phẩm tại một công ty. Bộ dữ liệu này gồm những cột như sau:
  + $APPID$: Thứ tự của các dòng (chạy từ 1 và đã được tráo thứ tự). Mục đích chỉ để biết được dòng đó là dòng nào.
  + $DISBURSALDATE$: Ngày bắt đầu cho vay.
  + $REPORT\_DATE$: Ngày nhận báo cáo tài chính.
  + $PRODUCT\_NAME$: Tên sản phẩm cho vay.
  + $LOANAMOUNT\_NOT\_INSURANCE$: Tiền vay chưa tính bảo hiểm.
  + $TERM\_TMP$: Khoảng cách thời gian từ $REPORT\_DATE$ đến $DISBURSALDATE$. (theo tháng)
  + $INSURANCE\_FEE$: Tiền phí bảo hiểm.
  + LA: Tổng tiền vay, bằng tổng $LOANAMOUNT\_NOT\_INSURANCE + INSURANCE\_FEE$.
  + EFF_RATE: Lãi suất theo năm.
  + $TERM$: thời gian vay (theo tháng)
  + $COUNT\_TERM\_RECEIPT$: Tổng số tháng đã trả tiền vay.
  + $SUM\_RECEIPT\_AMT$: Tổng tiền vay đã trả.
  + EMI: Tiền trả hằng tháng.
  + $GENDER$: Giới tính của người vay (F = Female = nữ, M = Male = nam).
  + AGE: Tuổi của người vay.
  + $NUMBER\_OF\_CHILDREN$: Số lượng con cái (để trống nghĩa là không có con).
  + $PERSONAL\_INCOME$: Tiền thu nhập cá nhân.
  + $ADDITIONAL\_INCOME$: Tiền thu nhập thêm.
  + $PERSONAL\_EXPENSE$: Phí sinh hoạt cá nhân.
  + $ADDITIONAL\_INCOME$: Tiền thu nhập của gia đình.
  + $FAMILY\_EXPENSE$: Phí sinh hoạt của gia đình.
  + $INCOME$: Tiền lợi nhuận, bằng ($PERSONAL\_INCOME$+$ADDITIONAL\_INCOME$ + $ADDITIONAL\_INCOME$)−($PERSONAL\_EXPENSE - FAMILY\_EXPENSE$).
  + $MARITAL\_STATUS$: Tình trạng hôn nhân.
  + $EDUCATION$: Trình độ học vấn.
  + $City$ và $ADDRESS\_CURRENT\_CITY$ và $PROVINCE\_CURRENT$: Nơi sinh sống.
  + (Quan trọng) $RESULT\_YES/NO$: Quyết định cho vay hoặc không (1 là có, 0 là không).
- Lưu ý: Nếu $COUNT\_TERM\_RECEIPT$ và $SUM\_RECEIPT\_AMT$ có giá trị là N/A, nghĩa là người đó có nợ xấu (vay nhưng không trả).

1.2 Tải tập dữ liệu

- Tải từ moodle.
- Tải từ link drive: [link]
- Dữ liệu bao gồm 1 file: $Data\_finance\_lab3.xlsx$

1.3 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Tải xuống và đọc được toàn bộ tập dữ liệu tài chính.
- Đọc dữ liệu từ file và in ra **5** dòng đầu tiên của tập dữ liệu.

2 Khám phá và tiền xử lý dữ liệu

2.1 Khám phá dữ liệu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Sử dụng các kỹ thuật thống kê để phân tích dữ liệu ban đầu. Từ đó chọn ra các cột dữ liệu cần thiết để sử dụng cho mô hình.
- Tiền xử lí dữ liệu: Dựa vào mô tả các cột ở phần một, một số cột sẽ có giá trị là chữ hoặc số hoặc có tính liên quan với nhau theo công thức. Vì vậy, nhóm cần một số bước xử lí dữ liệu trước khi huấn luyện mô hình sau khi đã chọn các cột dữ liệu cần sử dụng. Ví dụ, các cột có giá trị là chữ thì có thể chuyển đổi sang số, ... Lưu ý, giá trị quy đổi phải có ý nghĩa và có giải thích lý do chuyển đổi như vậy.
- Các nhóm được phép sử dụng thêm một số phương pháp khác Để gia tăng hiệu quả của mô hình nhưng nhóm cần cho biết phương pháp mình áp dụng là gì và cho biết mức độ cải thiện cụ thể gia tăng bao nhiêu.

3 Các yêu cầu về mô hình

3.1 Mô hình

Lab này không giới hạn cách làm cho việc phát hiện bất thường trong tài chính, nghĩa là đưa vào một dòng (có thể thiếu vài thông tin) mô hình nhóm cần xuất ra quyết định là (1 - bình thường, 0 - bất thường) . Lưu ý, tương tự như lab 1, nhóm chỉ được sử dụng những phương pháp **không có giám sát** (Unsupervised learning) và **không** được sử dụng deep learning (ngoại lệ cho những phương pháp đã được học trên lớp lý thuyết).

3.2 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Trình bày cấu trúc và cách thiết kế mô hình mình chọn một cách cụ thể, chi tiết từng bước tính toán từ đầu vào cho đến đầu ra.
- Trong mã nguồn, nếu nhóm sử dụng các tham số đặc biệt nào đó thì cần tìm hiểu và giải thích lý do tại sao chọn.
- Sau khi huấn luyện mô hình, nhóm cần thực hiện 2 hướng đi như sau:

+ Xem cột $RESULT\_YES/NO$ là giá trị đúng, nghĩa là giá trị tại cột này ở một dòng có thể hiểu là có dị thường hay không: giá trị 1 tương ứng với việc quyết định cho vay (bình thường), giá trị 0 tương ứng với quyết định không cho vay (bất thường). Dựa vào giá trị cột này, nhóm cần thực hiện các phép đo của mô hình mình để xem độ hiểu quả tới đâu. (Có khá nhiều độ đo nên cân nhắc cách chọn và giải thích được lý do chọn).
+ Xem cột $RESULT\_YES/NO$ là đầu ra của một mô hình bất kì và quyết định cho vay hay không dựa vào giá trị của cột này cũng có thể là một "bất thường" trong tài chính. Nhóm cần in ra được những dòng chứa dữ liệu bất thường dựa vào giá trị tại cột này. Gợi ý: Có nhiều hướng tiếp cận, nhóm có thể lấy cột này làm đầu vào của mô hình hoặc có thế không lấy vào đầu vào và so sánh sự khác nhau của đầu ra mô hình của nhóm và giá trị tại cột này,...

4 Các yêu cầu khác

- Ngôn ngữ sử dụng bắt buộc là Python, không được phép sử dụng ngôn ngữ khác. (nên sử dụng Jupiter Notebook).

- Không giới hạn thư viện được sử dụng trong Python.

- Các nhóm cần kiểm tra mã nguồn trước khi nộp. Nếu mã nguồn không chạy được mà không phải do nguyên nhân khách quan (thiếu thư viện, lỗi do thư viện gây ra, sử dụng thư viện sai phiên bản,...) thì sẽ bị 0 điểm đồ án.

- Bài nộp phải gồm có 3 phần:

  + Report: Chứa các file báo cáo.
  + Source: Chứa các file mã nguồn.
  + Presentation: Chứa các file dùng để thuyết trình.

- Trong các file nộp, nhóm cần ghi rõ thông tin về các thành viên gồm họ tên và MSSV. Riêng đối với mã nguồn, nhóm có thể ghi thông tin trên dưới dạng comment trong code của nhóm.

- Bài nộp sẽ được đặt trong thư mục có tên `MSSV01[_MSSV02[_MSSV03[...]]]` và được nén lại bằng định dạng ZIP với cùng tên như trên. Ví dụ đặt tên nhóm có 1 sinh viên là `MSSV01`, nhóm có 2 sinh viên là `MSSV01_MSSV02`.

- Nghiêm cấm các hành vi gian lận, không trung thực trong học tập như sao chép bài làm giữa các nhóm với nhau, sao chép bài làm của các nhóm khóa trước hoặc các nhóm lớp khác trường khác, nhờ người làm hộ. Nếu phát hiện các hành vi trên thì cả nhóm sẽ bị 0 điểm và xử lý theo quy định của Khoa và Trường.