ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

Đồ án chuyên ngành - CO4029

Báo cáo

KHAI PHÁ DỮ LIỆU NGÂN HÀNG - TUẦN 5

Giảng viên hướng dẫn: PGS.TS Trần Minh Quang
ThS. Đỗ Thanh Thái

Sinh viên thực hiện: 2014486 - Đậu Xuân Thành
2014848 - Nguyễn Xuân Thắng

TP. Hồ Chí Minh, 10/2023

Mục lục

Danh mục hình ảnh

1. Tổng quan

Báo cáo tuần 5 của nhóm tập trung vào giải quyết những vấn đề theo sự hướng dẫn của GVHD (Thầy Quang), bao gồm:

- Tìm kiếm dữ liệu về chủ đề ***Loan Approval Prediction*** và nghiên cứu, áp dụng những kỹ thuật khai phá dữ liệu lên đó.
- Bước đầu phân tích các yêu cầu chức năng, phi chức năng cho website của bộ công cụ.

2. Tìm kiếm & Thu thập dữ liệu (Data collection)

2.1. Tập dữ liệu 1 [1]

Mô tả: Tập dữ liệu được thu thập được các ngân hàng thu thập từ các khoản vay trước đây, nhằm mục đích xây dựng các mô hình dự đoán dựa trên các kỹ thuật khai phá dữ liệu, học máy. Để từ đó, phân loại các người đi vay xem họ có khả năng vỡ nợ hay không.

Thời gian thu thập dữ liệu: 2019
Độ lớn: hơn 140.000 dòng, với 34 thuộc tính.
Nguồn: Tập dữ liệu được giới thiệu bởi Kaggle - Một cộng đồng chia sẻ các nguồn dữ liệu đáng tin cậy, được nhiều bài báo về chủ đề dữ liệu, học máy sử dụng.

Các thuộc tính:

- *ID:* Mã số khoản vay.
- *year:* Năm dữ liệu được ghi nhận.
- *loan_limit:* Giới hạn khoản vay.
- *Gender:* Giới tính.
- *approv_in_adv:* Đã được phê duyệt trước.
- *loan_type:* Loại khoản vay.
- *loan_purpose:* Mục đích vay.
- *Credit_Worthiness:* Uy tín tín dụng.
- *open_credit:* Tín dụng mở. (Tín dụng xoay vòng)
- *business_or_commercial:* Kinh doanh hoặc thương mại
- *loan_amount:* Khoản vay.
- *rate_of_interest:* Lãi suất.
- *Interest_rate_spread:* Chênh lệch lãi suất (lãi suất cho vay - lãi suất tiền gửi).
- *Upfront_charges:* Lệ phí mà người đi vay phải trả trước khi khoản vay được chấp nhận.
- *term:* Kì hạn vay
- *Neg_ammortization:* Thoả thuận trả góp hàng tháng
- *interest_only:* Khoản vay theo lịch trình (chỉ cần trả lãi)
- *lump_sum_payment:* Khoản vay được chia thành nhiều đợt trả hay không
- *property_value:* Giá trị tài sản.
- *construction_type:*
- *occupancy_type:*
- *Secured_by:*
- *total_units:*
- *income:* Thu nhập của người vay.
- *credit_type:* Loại tổ chức đánh giá điểm tín dụng
- *Credit_Score:* Điểm tín dụng, được tính dựa trên lịch sử trả nợ và hồ sơ tín dụng của người vay.
- *co-applicant_credit_type:* Thang điểm đánh giá tín dụng của người đi vay cùng (nếu có).
- *age:* Tuổi

- *submission_of_application:*
- *LTV:* loan-to-value là tỉ lệ khoản tiền thế chấp trên giá trị thẩm định của tài sản
- *Region:* Khu vực
- *Security_Type:*
- *Status:* Kết quả đánh giá.
- *dtir1:*

2.2. Tập dữ liệu 2 [2]

Mô tả: Tập dữ liệu từ LendingClub.com, tập dữ liệu được thu thập từ các khoản đầu tư trong quá khứ để phân tích, đánh giá, giúp cho các nhà đầu tư có thể có xác suất đầu tư vào các đối tượng có khả năng hoàn vốn cao mình.

Thời gian thu thập dữ liệu: 2007 - 2010
Độ lớn: hơn 9.578 dòng, với 14 thuộc tính.
Nguồn: Tập dữ liệu được thu thập từ LendingClub.com - Nền tảng giúp kết nối những người cần vay tiền với những người có tiền để cho vay (nhà đầu tư). Đây là nền tảng cung cấp các dịch vụ tài chính có trụ sở tại Mỹ, có uy tín cao trong ngành tài chính, đầu tư.

Các thuộc tính:
- *credit.policy:* 1 nếu người dùng đáp ứng các tiêu chí đánh giá của LendingClub.com, và ngược lại là 0.
- *purpose:* Mục đích của khoản vay (nhận các giá trị: credit_card, debt_consolidation, educational, major_purchase, small_business, and all_other)
- *int.rate:* Lãi suất khoản vay, người đi vay được LendingClub.com đánh giá có mức rủi ro cao sẽ có mức lãi suất cao hơn
- *installment:* Số tiền trả góp hàng tháng nếu khoản vay được chấp nhận.
- *log.annual.inc:* Nhật ký về thu nhập hàng năm tự sao kê của người đi vay.
- *dti:* debt-to-income: Tỷ lệ nợ trên thu nhập của người đi vay.
- *fico:* Điểm tín dụng FICO của người đi vay.
- *days.with.cr.line:* Số ngày người đi vay tín dụng.
- *revol.bal:* Số dư xoay vòng của người đi vay (số tiền chưa thanh toán ở mỗi kỳ thanh toán tín dụng).
- *revol.util:* Tỉ lệ sử dụng hạn mức tín dụng xoay vòng của người cho vay.
- *inq.last.6mths:* Số lượng yêu cầu vay của người đi vay trong 6 tháng vừa qua.
- *delinq.2yrs:* Số lần người đi vay quá hạn hơn 30 ngày trong vòng 2 năm qua.
- *pub.rec:* Số lượng hồ sơ công khai của người vay.

Tài liệu tham khảo

[1] M. Y. H, "Loan default dataset," 2022. [Online]. Available: https://www.kaggle.com/datasets/yasserh/loan-default-dataset/data

[2] ItsSuru, "Loan data," 2021. [Online]. Available: https://www.kaggle.com/datasets/itssuru/loan-data/data