BÁO CÁO PHÂN TÍCH DỮ LIỆU

Hệ thống cho vay P2P - TIMA

Task 1: Overview & Data Understanding

Người thực hiện: Hoàng Thái Duy

Hà Nội, tháng 12 năm 2025

Mục lục

1 Task 1: Overview & Data Understanding **2**
1.1 1. Tìm hiểu nghiệp vụ & Xác định KPI 2
1.2 2. Phân tích dữ liệu & Mối quan hệ 4
1.3 3. Bộ câu hỏi phân tích (Analytical Questions) 7
1.4 4. Kế hoạch kiểm tra & làm sạch dữ liệu 8

1 Task 1: Overview & Data Understanding

1.1 1. Tìm hiểu nghiệp vụ & Xác định KPI

Trong bối cảnh TIMA hoạt động theo mô hình P2P Lending (Peer-to-Peer Lending), việc xác định và theo dõi các chỉ số đánh giá hiệu quả kinh doanh (KPIs) là vô cùng quan trọng để đảm bảo sự phát triển bền vững và kiểm soát rủi ro hiệu quả. Dưới đây là hệ thống KPIs được phân loại theo các cấp độ khác nhau:

KPIs cấp doanh nghiệp

Chỉ số tài chính tổng thể

- **Tổng doanh thu (Revenue)**: Tổng lãi + phí thu từ các khoản vay. Đây là chỉ số phản ánh quy mô và sức khỏe tài chính của TIMA, cho phép đánh giá khả năng tạo ra giá trị từ hoạt động cho vay.
- **Lợi nhuận ròng / EBITDA**: Đo lường mức sinh lời thực sự sau chi phí, giúp đánh giá hiệu quả hoạt động kinh doanh và khả năng tạo dòng tiền của doanh nghiệp.

KPIs Marketing / Sales

Nhóm KPIs này giúp đánh giá hiệu quả của các hoạt động marketing và bán hàng:

- **Số hồ sơ nộp (Applications)**: Đếm theo ngày/tuần/tháng dựa trên biến `application_date`. Chỉ số này phản ánh sức hút của sản phẩm và hiệu quả chiến dịch marketing.
- **Tỷ lệ phê duyệt (Approval Rate)**:

$$\text{Approval Rate} = \frac{\text{Approved}}{\text{Total Applications}} \times 100\%$$

  Thể hiện hiệu quả của quy trình thẩm định và tiêu chuẩn cho vay.

- **Tỷ lệ chuyển đổi sang giải ngân (Disbursement Conversion)**:

$$\text{Disbursement Conversion} = \frac{\text{Số khoản giải ngân}}{\text{Số hồ sơ được phê duyệt}} \times 100\%$$

- **Ticket Size trung bình**: Giá trị trung bình của `TienGiaiNgan`, phản ánh quy mô trung bình của các khoản vay.

KPIs Rủi ro & Danh mục

Đây là nhóm KPIs quan trọng nhất trong hoạt động cho vay, giúp kiểm soát và dự báo rủi ro:

- **Tỷ lệ nợ xấu (Bad Debt Rate)**:

$$\text{Bad Debt Rate} = \frac{\text{Số khoản vay có } \texttt{HasBadDebt=1}}{\text{Tổng số khoản vay}} \times 100\%$$

  Hoặc có thể tính dựa vào điều kiện `LongestOverdue` ≥ 90 ngày.

- **Tỷ lệ trả chậm (Late Payment Rate)**: Dựa trên biến `HasLatePayment`, đo lường tỷ lệ khách hàng có ít nhất một lần trả chậm trong lịch sử.

- **Exposure at Risk (EAD)**: Tổng giá trị `SoTienConLai` của các khoản vay đang trong tình trạng quá hạn, phản ánh tổng số tiền đang gặp rủi ro.

- **Time-to-default**: Thời gian trung bình từ `FromDate` (ngày giải ngân) đến thời điểm phát sinh quá hạn đầu tiên, giúp dự báo sớm rủi ro.

KPIs sản phẩm & địa bàn

- **ROI theo sản phẩm (Product ROI)**:

$$\text{Product ROI} = \frac{\text{Lợi nhuận từ sản phẩm}}{\text{Chi phí đầu tư vào sản phẩm}} \times 100\%$$

  Dánh giá hiệu quả kinh doanh theo từng loại sản phẩm (`ProductCreditName`).

- **Default Rate theo tỉnh/thành**: Phân tích tỷ lệ nợ xấu theo địa bàn địa lý (`CityName`), giúp xác định các khu vực có rủi ro cao.

KPIs khách hàng

- **Repeat Borrower Rate**:

$$\text{Repeat Rate} = \frac{\text{Số khách hàng vay } \geq 2 \text{ lần}}{\text{Tổng số khách hàng}} \times 100\%$$

  Tỉ lệ khách hàng vay nhiều lần (nhóm theo `ID`), phản ánh mức độ hài lòng và trung thành của khách hàng.

- **Customer Lifetime Value (CLTV)**: Ước tính giá trị mà một khách hàng mang lại trong suốt vòng đời quan hệ với TIMA, dựa trên dòng tiền tạo ra trừ đi chi phí phục vụ.

1.2 2. Phân tích dữ liệu & Mối quan hệ

Phân loại Dimensions và Measures

Để xây dựng một hệ thống phân tích dữ liệu hiệu quả, cần phân loại rõ ràng các trường dữ liệu thành hai nhóm chính:

Dimensions (Định tính)	**Measures (Định lượng)**
ID, LoanID, FullName	SoTienDKVayBanDau
Gender, Birthday	TienGiaiNgan
SoDienThoai	SoTienConLai
CityName, DistrictName, WardName	Salary
HouseHold	LongestOverdue
JobName, NameCompany	TS_CREDIT_SCORE_V2
ReceiveYourIncome	NumberOfLoans
ProductCreditName	
Trạng thái	
HasBadDebt, HasLatePayment	
application_date, FromDate, ToDate	

Bảng 1: Phân loại Dimensions và Measures trong dữ liệu TIMA

Giải thích:

- **Dimensions**: Các thuộc tính mô tả, phân loại dữ liệu, thường được sử dụng để lọc, nhóm và phân đoạn trong phân tích.
- **Measures**: Các giá trị số có thể tính toán (tổng, trung bình, min, max), là đối tượng chính của các phép phân tích định lượng.

Sơ đồ mối quan hệ giả định

Dưới đây là sơ đồ luồng dữ liệu và mối quan hệ giữa các nhóm biến trong quy trình cho vay:

Data Flow Diagram

```
    [Customer Profile]
- Age, Gender
- JobName, Salary
- NumberOfLoans




    [Credit Signals]
- TS_CREDIT_SCORE_V2
- HasBadDebt
- HasLatePayment




  [Underwriting Decision]
- Approval Status
- TienGiaiNgan
- Ticket Size




    [Loan Performance]
- LongestOverdue
- SoTienConLai
- Default Status
```

Các giả thuyết nghiệp vụ chính

Business Hypotheses

1. **Giả thuyết về thu nhập**: Khách hàng có `Salary` cao có xu hướng được giải ngân số tiền cao hơn (`TienGiaiNgan`) và có rủi ro quá hạn thấp hơn do khả năng thanh toán tốt.
2. **Giả thuyết về điểm tín dụng**: Khách hàng có `TS_CREDIT_SCORE_V2` thấp có xác suất rơi vào trạng thái nợ xấu (default) cao hơn đáng kể.
3. **Giả thuyết về đa dạng vay**: Khách hàng có `NumberOfLoans` cao (vay từ nhiều nguồn) có rủi ro vay chéo (over-indebtedness) và khả năng không trả được nợ cao hơn.
4. **Giả thuyết về tài sản đảm bảo**: Các sản phẩm cho vay có tài sản đảm bảo (như cầm cố xe máy/ô tô) thường có tỷ lệ nợ xấu thấp hơn so với các khoản vay tín chấp do có thể thu hồi tài sản khi khách hàng không trả được nợ.

1.3 3. Bộ câu hỏi phân tích (Analytical Questions)

Để đảm bảo phân tích dữ liệu toàn diện và có chiều sâu, chúng ta xây dựng hai nhóm câu hỏi chính theo mô hình phân tích mô tả và chẩn đoán:

Descriptive Analytics (Phân tích mô tả)

Nhóm câu hỏi này tập trung vào việc mô tả hiện trạng và xu hướng của dữ liệu:

1. **Xu hướng thời gian**: Xu hướng số hồ sơ nộp và tổng giá trị giải ngân theo tháng/quý là gì? Có tính chu kỳ hay mùa vụ nào không?
2. **Phân tích sản phẩm**: Ticket Size trung bình của từng loại sản phẩm (`ProductCreditName`) là bao nhiêu? Sản phẩm nào chiếm tỷ trọng lớn nhất?
3. **Đánh giá rủi ro**: Tỷ lệ nợ xấu của từng loại sản phẩm là bao nhiêu? Sản phẩm nào có rủi ro cao nhất?
4. **Phân phối điểm tín dụng**: Phân phối của `TS_CREDIT_SCORE_V2` như thế nào theo các trạng thái hồ sơ khác nhau (phê duyệt, từ chối, nợ xấu)?
5. **Hành vi khách hàng**: Tỉ lệ khách hàng vay lại (Repeat Borrower Rate) là bao nhiêu? Họ có đặc điểm gì khác biệt so với khách hàng mới?

Diagnostic Analytics (Phân tích chẩn đoán)

Nhóm câu hỏi này tìm hiểu nguyên nhân sâu xa đằng sau các hiện tượng được quan sát:

1. **Phân tích nợ xấu theo sản phẩm**: Vì sao tỷ lệ nợ xấu của một số sản phẩm như "Vay theo Sim" lại cao hơn đáng kể so với các sản phẩm khác? Có phải do đặc điểm khách hàng hay chính sách sản phẩm?
2. **Nhân tố ảnh hưởng đến quá hạn**: Những biến số nào (thu nhập, điểm tín dụng, số khoản vay hiện có, địa lý) có ảnh hưởng mạnh nhất đến `LongestOverdue`? Có thể xây dựng mô hình dự báo không?
3. **Phân tích chuyển đổi**: Vì sao tỷ lệ chuyển đổi từ hồ sơ được phê duyệt sang giải ngân thực tế lại thấp tại một số tỉnh/thành phố cụ thể? Có phải do vấn đề logistics, văn hóa hay cạnh tranh?
4. **Segmentation rủi ro-giá trị**: Nhóm khách hàng nào có Ticket Size lớn (mang lại doanh thu cao) nhưng đồng thời có rủi ro cao? Làm thế nào để cân bằng giữa tăng trưởng và kiểm soát rủi ro?
5. **Hiệu quả quy trình**: Nguyên nhân nào khiến thời gian từ phê duyệt đến giải ngân (time-to-disbursement) bị kéo dài? Có điểm nghẽn nào trong quy trình không?

1.4 4. Kế hoạch kiểm tra & làm sạch dữ liệu

Chất lượng dữ liệu là nền tảng cho mọi phân tích chính xác. Dưới đây là kế hoạch chi tiết để đảm bảo tính toàn vẹn và độ tin cậy của dữ liệu:

Các vấn đề chất lượng dữ liệu cần kiểm tra

Vấn đề	Mô tả chi tiết
Missing values	Kiểm tra các trường quan trọng: `Salary`, `LongestOverdue`, `TS_CREDIT_SCORE_V2`, các trường ngày tháng. Đánh giá mức độ thiếu và pattern.
Sai kiểu dữ liệu	`Salary` đang ở dạng text (có ký tự phân cách), các trường ngày ở dạng string cần convert sang datetime.
Trùng lặp	Kiểm tra trùng `LoanID` (khóa chính), phát hiện các bản ghi duplicate hoàn toàn hoặc một phần.
Giá trị bất thường	• `SoTienConLai` $>$ `TienGiaiNgan` • `TienGiaiNgan` hoặc `Salary` âm • `LongestOverdue` âm hoặc vượt quá tuổi khoản vay • Điểm tín dụng ngoài khoảng hợp lệ
Ngày không hợp lệ	`ToDate` $<$ `FromDate`, `application_date` trong tương lai, ngày sinh không hợp lý (tuổi $<$ 18 hoặc $>$ 100).
Chuẩn hóa danh mục	Thống nhất cách viết cho `ProductCreditName`, `CityName`, `JobName` (viết hoa, dấu cách, ký tự đặc biệt).
PII (Thông tin cá nhân)	Cần ẩn hoặc mã hóa `FullName`, `SoDienThoai` khi xuất báo cáo hoặc chia sẻ dữ liệu.

Bảng 2: Danh sách các vấn đề chất lượng dữ liệu

Quy trình xử lý dữ liệu

Data Cleaning Pipeline

Bước 1: Làm sạch dữ liệu số
- Làm sạch `Salary` bằng cách loại bỏ các ký tự không phải số (dấu phân cách, ký tự đặc biệt).
- Chuyển đổi sang kiểu numeric và xử lý các giá trị không hợp lệ.

Bước 2: Chuẩn hóa biến nhị phân
- Chuyển `HasBadDebt`, `HasLatePayment` sang kiểu boolean hoặc 0/1.
- Kiểm tra tính nhất quán giữa các biến liên quan.

Bước 3: Tạo Feature Engineering
- `Age`: Tính từ `Birthday` đến thời điểm hiện tại.
- `AgeGroup`: Phân nhóm tuổi (18-25, 26-35, 36-45, 46-55, 56+).
- `DaysActive`: Số ngày từ `FromDate` đến `ToDate`.
- `TicketBucket`: Phân nhóm Ticket Size (<5M, 5-10M, 10-20M, 20M+).
- `DebtToIncome`: Tỷ lệ nợ trên thu nhập.

Bước 4: Xử lý dữ liệu sai
- Loại bỏ các bản ghi có lỗi nghiêm trọng không thể sửa.
- Đánh dấu (Flag) các bản ghi nghi ngờ để xem xét thêm.
- Ghi log chi tiết các thay đổi để truy vết sau này.

Checklist kiểm tra sau làm sạch

1. ✓Không còn giá trị missing ở các trường bắt buộc
2. ✓Tất cả các trường số đã đúng kiểu dữ liệu
3. ✓Không còn bản ghi trùng lặp
4. ✓Các giá trị đều nằm trong khoảng hợp lệ
5. ✓Các trường ngày tháng đã được validate
6. ✓Danh mục đã được chuẩn hóa thống nhất
7. ✓PII đã được bảo mật phù hợp

Kết luận

Báo cáo Task 1 đã hoàn thành việc khảo sát tổng quan và hiểu biết sâu về dữ liệu của hệ thống P2P Lending TIMA. Các nội dung chính bao gồm:

- Xác định hệ thống KPIs đa chiều từ cấp doanh nghiệp đến cấp khách hàng
- Phân tích cấu trúc dữ liệu, mối quan hệ giữa các biến và các giả thuyết nghiệp vụ
- Xây dựng bộ câu hỏi phân tích mô tả và chẩn đoán
- Lập kế hoạch chi tiết cho việc kiểm tra và làm sạch dữ liệu

Những kết quả này sẽ là nền tảng vững chắc cho các giai đoạn phân tích sâu và xây dựng mô hình trong các task tiếp theo.