|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **BÁO CÁO THỰC TẬP TỐT NGHIỆP**  **DỰ ĐOÁN KHẢ NĂNG TRẢ NỢ CỦA KHÁCH HÀNG BẰNG MÔ HÌNH HỌC MÁY**  **Giảng viên hướng dẫn: TRẦN CHÂU THANH THIỆN**  **Sinh viên thực hiện : ĐẶNG THANH PHÚC**  **MSSV : 2100009466**  **Đơn vị thực tập : Viện Cơ Học Và Tin Học Ứng Dụng Khóa : 2021-2025**  **Tp.HCM, tháng 11 năm 2024** |

|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **BÁO CÁO THỰC TẬP TỐT NGHIỆP**  **DỰ ĐOÁN KHẢ NĂNG TRẢ NỢ CỦA KHÁCH HÀNG BẰNG MÔ HÌNH HỌC MÁY**  **Giảng viên hướng dẫn : TRẦN CHÂU THANH THIỆN**  **Sinh viên thực hiện : ĐẶNG THANH PHÚC**  **MSSV : 2100009466**  **Đơn vị thực tập : Viện Cơ Học Và Tin Học Ứng Dụng**  **Khóa : 2021-2025**  **Tp.HCM, tháng 11 năm 2024** |

# 

# LỜI CẢM ƠN

Trong quá trình hoàn thiện và nghiên cứu dự án của kì thực tập này, em xin gửi lời cảm ơn chân thành đến thầy Hồ Khôi và thầy Trần Châu Thanh Thiện đã dành thời gian và công sức giảng dạy và chia sẻ kiến thức đến với em. Sự giúp đỡ của thầy và các bạn đã giúp em hoàn thành được dự án này một cách tốt nhất. Em cũng muốn bày tỏ lòng biết ơn đến quý thầy cô, bạn bè đã luôn đồng hành, chia sẻ kiến thức và tạo điều kiện tạo nên môi trường học tập tích cực.

Những kiến thức quý báu và sự hỗ trợ của quý thầy cô đã giúp em phát triển kỹ năng nghiên cứu và rèn luyện tư duy logic cũng như củng cố cho mình một nền tảng để tiếp tục phát triển thêm trong tương lai. Sự hỗ trợ đó đã trở thành nguồn động viên to lớn với em trong những giai đoạn khó khăn nhất. Qua đó cũng đã tạo ra một môi trường học tập và làm việc hiệu quả, tích cực cho sinh viên như em có thể tiếp thu kiến thức và phát triển một cách tốt nhất.

Em cũng xin chân thành cảm ơn nhóm 2 vì sự hợp tác, hỗ trợ và những đóng góp thiết thực trong từng giai đoạn thực hiện dự án. Chính sự đồng hành của các bạn đã tạo nên tinh thần làm việc nhóm hiệu quả và sự gắn kết chặt chẽ, giúp chúng ta vượt qua các khó khăn.

Ngoài ra em xin được gửi lời cảm ơn đến với VIỆN CƠ HỌC VÀ TIN HỌC ỨNG DỤNG đã tạo điều kiện cho em thực tập tại đây, giúp em có được kinh nghiệm và trải nghiệm thực tế công việc, tích lũy thêm được nhiều kiến thức về chuyên ngành em đang theo đuổi.

Em xin chân thành cảm ơn.

# LỜI MỞ ĐẦU

Trong kỷ nguyên Cách mạng Công nghiệp 4.0, trí tuệ nhân tạo và các công nghệ phân tích dữ liệu ngày càng chứng minh vai trò quan trọng trong việc hỗ trợ con người đưa ra những quyết định dựa trên cơ sở dữ liệu. Logistic Regression, một trong những thuật toán học máy phổ biến nhất, đã trở thành một công cụ mạnh mẽ trong các bài toán phân loại, đặc biệt trong lĩnh vực tài chính, y tế, và tiếp thị.

Logistic Regression không chỉ đơn giản trong lý thuyết mà còn hiệu quả trong thực tế khi làm việc với dữ liệu nhị phân hoặc đa phân loại. Thuật toán này hoạt động dựa trên việc ước lượng xác suất một sự kiện xảy ra, chẳng hạn như khả năng một khách hàng hoàn trả khoản vay, hay một giao dịch là gian lận. Đặc biệt, với bản chất tuyến tính trong việc phân tách không gian đầu vào, Logistic Regression có thể xử lý tốt các vấn đề với dữ liệu quy mô lớn mà không đòi hỏi tài nguyên tính toán quá nhiều.

Logistic Regression là một công cụ mạnh mẽ, linh hoạt và dễ triển khai trong các bài toán phân loại. Nhờ khả năng xử lý dữ liệu và tính toán hiệu quả, mô hình này không chỉ giúp tối ưu hóa kết quả dự đoán mà còn mang lại giá trị thực tiễn trong nhiều lĩnh vực. Với sự phát triển của dữ liệu lớn và công nghệ máy học, Logistic Regression sẽ tiếp tục đóng vai trò quan trọng trong việc hỗ trợ các nhà phân tích và doanh nghiệp đưa ra quyết định một cách chính xác, nhanh chóng và khoa học.

# NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

*TPHCM, Ngày …… tháng …… năm 2023*

**Giáo viên hướng dẫn**

(Ký tên, đóng dấu)

# NHẬN XÉT CỦA CƠ QUAN THỰC TẬP

*TPHCM, Ngày …… tháng …… năm 2023*

**Xác nhận của đơn vị thực tập**

(Ký tên, đóng dấu)

**MỤC LỤC**

[LỜI CẢM ƠN i](#_Toc183772056)

[LỜI MỞ ĐẦU ii](#_Toc183772057)

[NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN iii](#_Toc183772058)

[NHẬN XÉT CỦA CƠ QUAN THỰC TẬP iv](#_Toc183772059)

[DANH MỤC CÁC BẢNG HÌNH v](#_Toc183772060)

[I. GIỚI THIỆU TỔNG QUAN VỀ VIỆN CƠ HỌC VÀ TIN HỌC ỨNG DỤNG 1](#_Toc183772061)

[1. Khái quát về Viện Cơ học và Tin học ứng dụng 1](#_Toc183772062)

[2. Quá trình hình thành và phát triển 3](#_Toc183772063)

[3. Cơ cấu tổ chức 3](#_Toc183772064)

[4. Các hoạt động thường xuyên của đơn vị 4](#_Toc183772065)

[5. Các thành tích nổi bật 5](#_Toc183772066)

[II. NỘI DUNG THỰC TẬP VÀ CƠ SỞ TRIỂN KHAI DỰ ÁN 7](#_Toc183772067)

[1. Thông tin quá trình tham gia thực tập 7](#_Toc183772068)

[2. Giới thiệu đề tài Dự đoán khả năng trả nợ của khách hàng bằng học máy 7](#_Toc183772069)

[3. Cơ sở lý thuyết 10](#_Toc183772070)

[III. XÂY DỰNG MÔ HÌNH 26](#_Toc183772071)

[1. Tổng quan phương pháp xây dựng 26](#_Toc183772072)

[2. Quy trình xây dựng mô hình 30](#_Toc183772073)

[IV. THỰC NGHIỆM 34](#_Toc183772074)

[1. Xây dựng thực nghiệm 34](#_Toc183772075)

[3. Kết quả thực nghiệm 40](#_Toc183772076)

[V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 41](#_Toc183772077)

[1. Kết luận 41](#_Toc183772078)

[2. Hướng phát triển 41](#_Toc183772079)

[TÀI LIỆU THAM KHẢO 42](#_Toc183772080)

# DANH MỤC CÁC BẢNG HÌNH

[Hình 1: Viện cơ học và tin học ứng dụng 1](#_Toc183771654)

[Hình 2: Armin Ronacher 11](#_Toc183771655)

[Hình 3: Các tính năng của Flask Framework 11](#_Toc183771656)

[Hình 4: Tại sao chọn flask 13](#_Toc183771657)

[Hình 5: Phương trình quan hệ 17](#_Toc183771658)

[Hình 6: Phương trình hồi quy 18](#_Toc183771659)

[Hình 7: Biểu đồ quá trình huấn luyện 29](#_Toc183771660)

[Hình 8: Trực quan hóa phân phối điểm FICO 30](#_Toc183771661)

[Hình 9: Trực quan hóa phân phối mục đích vay 31](#_Toc183771662)

[Hình 10: Trực quan hóa mối quan hệ giữa điểm FICO và trạng thái trả nợ 31](#_Toc183771663)

[Hình 11: Trực quan hóa ma trận tương quan 32](#_Toc183771664)

[Hình 12: Kết quả huấn luyện model Logistic Regression 34](#_Toc183771665)

[Hình 13: Kết quả huấn luyện model Random Forests 35](#_Toc183771666)

[Hình 14: Giao diện cải tiến hoàn chỉnh 36](#_Toc183771667)

[Hình 15: Nhập thông tin 36](#_Toc183771668)

[Hình 16: Các thông số cần nhập và thông tin đã nhập 36](#_Toc183771669)

[Hình 17: Kết quả dự đoán 37](#_Toc183771670)

[Hình 18: Hiển thị lên lịch sử 37](#_Toc183771671)

[Hình 19: Xem lịch sử chi tiết 38](#_Toc183771672)

[Hình 20: Giao diện đăng nhập admin 39](#_Toc183771673)

[Hình 21: Tất cả lịch sử giao dịch 39](#_Toc183771674)

[Hình 22: Xem chi tiết trong phần lịch sử giao dịch 40](#_Toc183771675)

# I. GIỚI THIỆU TỔNG QUAN VỀ VIỆN CƠ HỌC VÀ TIN HỌC ỨNG DỤNG

## 1. Khái quát về Viện Cơ học và Tin học ứng dụng

### 1.1 Thông tin chung

A building with glass windows

Description automatically generated

Hình 1: Viện cơ học và tin học ứng dụng

Tên đơn vị: Viện Cơ học và Tin học ứng dụng

Tên quốc tế: IAIM

Cơ quan chủ quản: Viện hàn lâm Khoa Học và Công Nghệ Việt Nam

Loại hình tổ chức: Viện, trung tâm NC&PT

Lĩnh vực nghiên cứu: Khoa học kỹ thuật và công nghệ

Loại hình hoạt động KH&CN chính: Nghiên cứu ứng dụng

Địa chỉ: 291 Điện Biên Phủ, Phường 7, Quận 3, TP. Hồ Chí Minh

Mã số thuế: 0301454159

Điện thoại: 0839307876

Fax: (+84)(28) 3930.8300

### 1.2 Chức năng và nhiệm vụ

* Nghiên cứu những vấn đề khoa học cơ bản làm cơ sở cho việc phát triển các hướng công nghệ mới và ứng dụng trong lĩnh vực: cơ học biển, dầu khí, máy móc, công trình, vật liệu xây dựng, chế tạo máy, tự động hóa, robot.
* Nghiên cứu ứng dụng và chuyển giao công nghệ mới trong lĩnh vực cơ học và công nghệ thông tin phục vụ quản lý và xử lý môi trường nước, đất, không khí.
* Nghiên cứu, khảo sát, xác định các chỉ tiêu cơ – lý của đất, đá, vật liệu, các điều kiện địa cơ học phục vụ thiết kế, thi công các công trình xây dựng.
* Triển khai các kết quả nghiên cứu và ứng dụng các công nghệ mới để thiết kế, chế tạo thiết bị phục vụ các ngành công nghiệp dược, y tế, thực phẩm, hàng tiêu dùng, xây dựng, công nghệ môi trường và các ngành công nghiệp khác.
* Nghiên cứu những vấn đề cơ bản của công nghệ thông tin và khoa học tính toán.
* Ứng dụng và triển khai công nghệ thông tin trong các lĩnh vực kinh tế - xã hội, khoa học, kỹ thuật.
* Sản xuất thử nghiệm trong lĩnh vực cơ học ứng dụng và công nghệ thông tin.
* Thực hiện các dịch vụ khoa học công nghệ trong lĩnh vực cơ học và công nghệ thông tin (đo đạc, kiểm định, tư vấn, thông tin, hội thảo khoa học,…).
* Tham gia với các ngành, các địa phương trong việc xét duyệt luận chứng kinh tế kỹ thuật và chuyển giao công nghệ trong lĩnh vực cơ học và công nghệ thông tin.
* Tham gia đào tạo cán bộ khoa học có trình độ cao trong lĩnh vực cơ học và công nghệ thông tin.
* Thực hiện hợp tác quốc tế trong nghiên cứu khoa học, đào tạo và phát triển công nghệ về cơ học và công nghệ thông tin.

**Những định hướng nghiên cứu - triển khai của Viện:**

* Cơ khí và Tự động hóa
* Cơ học Tính toán, Cơ học Công trình và Chẩn đoán kỹ thuật
* Cơ học Thủy, khí – Cơ học sông, biển và Môi trường
* Công nghệ thông tin

## 2. Quá trình hình thành và phát triển

Viện Cơ học và Tin học ứng dụng là một tổ chức quốc gia đa ngành, được thành lập bởi Chính phủ từ năm 1984. Trải qua nhiều năm phát triển, Viện đã đầu tư mạnh mẽ vào các trang thiết bị chuyên ngành tiên tiến, đặt mình vào vị trí hàng đầu trong lĩnh vực này. Như một đơn vị thuộc Chính phủ, nhiệm vụ chính của Viện là tiến hành nghiên cứu cơ bản trong lĩnh vực khoa học tự nhiên và phát triển công nghệ.

Viện đóng vai trò quan trọng trong việc cung cấp luận cứ khoa học để hỗ trợ quản lý trong lĩnh vực khoa học và công nghệ, cũng như đóng góp vào việc xây dựng chính sách, chiến lược, và quy hoạch phát triển kinh tế-xã hội. Ngoài ra, Viện còn chịu trách nhiệm đào tạo nhân lực chất lượng cao với trình độ chuyên môn theo quy định của pháp luật.

## 3. Cơ cấu tổ chức

### 3.1 Lực lượng cán bộ

* Tổng số CBVC: 74
* Số biên chế: 66
* Số hợp đồng: 08
* Phó giáo sư: 04
* Tiến sĩ: 07
* Thạc sĩ: 17
* Kỹ sư/Cử nhân: 40
* Khác: 10

### 3.2 Các phòng chuyên môn

* Phòng Cơ học máy
* Phòng Địa cơ học - Công trình ngầm và Chuẩn đoán kỹ thuật
* Phòng Cơ học biển và Môi trường
* Phòng Cơ học tính toán và Công trình
* Phòng Tự động hóa và rô bốt
* Phòng Hệ thống thông tin quản lý
* Phòng Hệ thống mạng
* Phòng Công nghệ tính toán và Công nghệ tri thức
* Phòng Nghiên cứu, ứng dụng Hệ thống thông tin

### 3.3 Các đơn vị quản lý nghiệp vụ

* Phòng chức năng: Phòng Quản lý tổng hợp
* Trung tâm Công nghệ
* Trung tâm Ứng dụng và công nghệ môi trường
* Trung tâm Nghiên cứu ứng dụng và tư vấn khoa học môi trường
* Trung tâm Quan trắc và ứng dụng công nghệ môi trường
* Xưởng Cơ nhiệt và thiết bị môi trường
* Xưởng Cơ khí chính xác
* Xưởng Cơ khí chế tạo máy
* Xưởng Vật liệu và máy công nghệ
* Xưởng Chế tạo thiết bị sản xuất vật liệu xây dựng
* Xưởng Chế tạo thiết bị tự động công nghiệp
* Xưởng chế tạo phụ tùng và công nghệ polymer

## 4. Các hoạt động thường xuyên của đơn vị

* Thực hiện đề tài nghiên cứu khoa học, dự án các cấp
* Triển khai các kết quả nghiên cứu vào thực tiễn
* Nghiên cứu ứng dụng và nghiên cứu phát triển công nghệ
* Thực hiện các dịch vụ KHKT: Thiết kế, chế tạo các thiết bị phục vụ cho các ngành công nghiệp, tư vấn và thi công các công trình về môi trường; thông tin, tư vấn, đo đạt, kiểm định, hiệu chỉnh, chuyển giao công nghệ, đào tạo nâng cao nghiệp vụ chuyên môn; thực hiện các dịch vụ KH&CN khác trong lĩnh vực cơ học và Tin học
* Thực hiện công tác đào tạo sau đại học
* Thực hiện nhiệm vụ hợp tác quốc tế

## 5. Các thành tích nổi bật

### 5.1 Lĩnh vực nghiên cứu cơ bản

Nghiên cứu các bài toán về ổn định, các bài toán động lực của các cơ hệ chịu tải phức tạp (hệ động lực phi tuyến và hỗn độn, hệ động lực ngẫu nhiên, ổn định các hệ máy, chi tiết như ổ trục-vòng bi, cân bằng rotor,...).

Nghiên cứu ổn định của công trình ngầm trong đất yếu.

Thiết lập mô hình toán-cơ cho các bài toán thủy động lực nước mặt, nước ngầm; lập trình và mô phỏng các bài toán lan truyền chất,...

Nghiên cứu các phương pháp số giải các bài toán biên trong cơ học và kỹ thuật với việc sử dụng công nghệ thông tin trợ giúp.

Lập trình và tính toán song song và phân tán, tính toán hình thức. Phát triển các hệ điều hành, ngôn ngữ lập trình,...

Nghiên cứu các chỉ tiêu cơ lý của vật liệu; xác định các tham số của cơ hệ; lý thuyết chẩn đoán kỹ thuật đối với công trình, thiết bị, máy móc,...

### 5.2 Các lĩnh vực khoa học công nghệ

Nghiên cứu tương tác biển với công trình là cơ sở khoa học cho việc xây dựng Quy phạm công trình biển.

Bài toán về trường độ thấm tại các mỏ dầu phục vụ quy hoạch khai thác hiệu quả.

Nghiên cứu tính toán ổn định các đập thủy điện chịu tải phức tạp.

Nghiên cứu tính toán lún sụp mặt đất; sạt lở đường bờ, mái dốc,...

Nghiên cứu thiết lập các phần mềm, tính toán các bài toán động học sông, biển và lan truyền chất ứng dụng trong dự báo ô nhiễm môi trường nước, không khí, thiết kế đê đập; các phần mềm ứng dụng trong chẩn đoán kỹ thuật, trong kinh tế, trong khám chữa bệnh từ xa; trong quản lý nhà nước về khoa học-công nghệ, tài nguyên-môi trường, giao thông-vận tải,

Cơ sở tính toán lũ cho Đồng bằng SCL; Dự báo ô nhiễm vùng hồ, cửa sông; Xây dựng cơ sở dữ liệu biển đông trên cơ sở thu ảnh vệ tinh nhỏ phục vụ dự đoán lũ lụt đồng bằng sông Cửu long, về nước dâng, về thảm thực vật, biến đổi đường bờ sông biển.

### 5.3 Nghiên cứu ứng dụng triển khai công nghệ

Thiết kế, chế tạo và sản xuất thử nghiệm các thiết bị và dây chuyền công nghệ đòi hỏi kỹ thuật và mức độ tự động hóa cao như: Phục hồi nhà máy sản xuất sữa Dielac Đồng Nai, dây chuyền sản xuất thuốc lá Khánh Hòa, dây chuyền sản xuất giầy da Hà Nội, dây chuyền sản xuất tấm lợp,…

Thiết kế, chế tạo các dây chuyền công nghệ tự động, các thiết bị cân kỹ thuật số; các thiết bị trong y khoa để chẩn đoán bệnh; các thiết bị đo tự động(cân băng tải, cân tàu hỏa-toa xe, cân ô tô); các thiết bị phân tích, xác định độ tro; các thiết bị cắt tia nước; thiết bị CNC, thiết bị mạ bốc bay chân không, thiết bị sấy gỗ nhanh siêu tần,...

Thiết kế chế tạo lắp đặt các thiết bị áp lực, thiết bị nâng, thiết bị nghiền sàng,… cho các nhà máy dệt, may, chế biến thực phẩm, sản xuất xi măng, đá, sản xuất gạch, vật liệu xây dựng,…

Thiết kế chế tạo và lắp đặt các thiết bị cơ khí chính xác cho ngành y tế, ngành dược, cho các nhà máy sản xuất thực phẩm trong cả nước,...

Thiết kế, chế tạo, lắp đặt thiết bị bơm hút chân không, máy nén, ép, cắt.

# II. NỘI DUNG THỰC TẬP VÀ CƠ SỞ TRIỂN KHAI DỰ ÁN

## 1. Thông tin quá trình tham gia thực tập

### 1.1 Vị trí thực tập

Vị trí: Kỹ sư trí tuệ nhân tạo – AI Engineer

Thời gian thực tập: 19/09/2024 – 30/11/2024

Người cố vấn tại đơn vị: Hồ Khôi

Hình thức làm việc:

### 1.2 Các yêu cầu thực tập

**Yêu cầu công việc bao gồm:**

* Đề tài thực hiện: Dự đoán khả năng trả nợ của khách hàng bằng mô hình học máy
* Lên kế hoạch cho toàn bộ quá trình thực tập
* Một bản báo cáo thực tập (Word)
* Một bản trình bày công việc thực tập (PowerPoint)
* Hoàn thành mã nguồn hoàn chỉnh của dự án

**Yêu cầu đề tài:**

* + Đề tài thuộc lĩnh vực Data hoặc AI
  + Có ứng dụng cao trong thực tế
  + Sử dụng các mô hình machine Learning hoặc Deeplearning để thực hiện công việc như dự đoán, phân loại, đưa ra quyết định. Tối thiểu 2 mô hình.
  + Xây dựng đầy đủ mô hình
  + Có giao diện cho người dùng (Ứng dụng hoặc website)

## Giới thiệu đề tài Dự đoán khả năng trả nợ của khách hàng bằng học máy

Trong lĩnh vực ngân hàng, tài chính, việc dự đoán, đánh giá khả năng trả nợ của khác hàng là một trong những yếu tố quyết định để giúp các tổ chức vay tín dụng và ngân hàng giảm đi những rủi ro khi cung cấp các thông tin và khoản vay. Thị trường vay nợ là một thị trường không thể không có trên các lĩnh vực kinh doanh, marketing,…từ đó tạo ra các khoản sinh lãi cho người cho vay. Để cải thiện được độ chính xác cao trong các mô hình dự đoán, các phương pháp học máy ( machine learning ) đã được sinh ra và áp dụng rộng rãi trên toàn thế giới. Những mô hình không chỉ phân tích thông tin tài chính mà còn xem xét các yếu tố như lịch sử tín dụng, thu nhập và các yếu tố khác để đưa ra một dự đoán chính xác hơn.

Chính vì thế mà việc nghiên cứu để xây dựng một hệ thống hay một ứng dụng dùng để áp dụng cho công nghệ trong lĩnh vực “ Machine learning “ và “ Deep learning “ đào tạo ra các mô hình dự đoán về khả năng trả nợ của khách hàng hỗ trợ đưa ra các dự đoán cho người sử dụng sẽ là một công cụ vô cùng quan trọng và cần thiết với thị trường hiện nay.

Mô hình Logistic Regression hay còn được gọi là “ Hồi quy Logistic “ là một mô hình học máy được sử dụng rộng rãi trong các bài toán loại nhị phân. Trong bài toán dự đoán khả năng trả nợ của khách hàng, Logistic Regression sẽ phân tích các yếu tố như thu nhập, lịch sử tín dụng và các đặc điểm khác để đưa ra xác suất khách hàng có khả năng trả nợ hay không.

Mô hình XGBoost có tên đầy đủ là eXtreme Gradient Boosting, là một thuật toán học máy mạnh mẽ thường được sử dụng trong các bài toán phân loại và hồi quy. XGBoost thuộc nhóm các thuật toán “ ensemble “, được gọi là mô hình kết hợp, mà cụ thể là phương phác Boosting giúp xây dựng một mô hình dự đoán mạnh bằng cách kết hợp nhiều mô hình yếu. Trong bài toán dự đoán khả năng trả nợ của khách hàng, XGBoost có thể phân tích các đặc điểm của khách hàng như thu nhập, lịch sử tín dụng, nợ quá hạn và các yếu tố khác để dự đoán xác suất trả nợ. XGBoost đưa ra các dự đoán chính xác, mức độ hiệu quả hơn so với Logistic.

Đề tài đặt ra mục tiêu là sẽ xây dựng và phát triển một hệ thống dự đoán khả năng trả nợ của khách hàng bằng học máy có độ chính xác cao dựa trên hai mô hình. Từ đó có thể giúp người sử dụng có thể nắm bắt được khả năng khách hàng có trả nợ hay không.

Trong phần kế tiếp, đề tài sẽ được trình bày về các nền tảng lý thuyết, phương pháp thực hiện, quy trình thực nghiệm cùng với kết quả đánh giá độ chính xác. Mục tiêu chung là nghiên cứu này sẽ góp phần đẩy mạnh khai thác nhiều hơn các mô hình dự đoán cụ thể là XGBoost, giúp chúng trở nên thông minh và hiệu quả hơn trong việc hỗ trợ người dùng nắm bắt được khả năng trả nợ của khách hàng

### Lý do chọn đề tài

Lý do đằng sau việc phát triển mô hình Logistic Regression và XGBoost để dự đoán xuất phát từ tầm quan trọng của việc đưa ra dự đoán là có thể trả nợ hay không. Trong thị trường hiện nay, số khách hàng vay vốn không thể được nói là ít, thậm chí rất nhiều trên toàn thế giới. Việc đánh giá khả năng trả nợ của khách hàng là nhiệm vụ quan trọng đối với các tổ chức cho vay tín dụng và ngân hàng. Yếu tố này sẽ ảnh hưởng tới quyết định đúng đắn trong việc cho vay không chỉ giúp bảo vệ lợi ích của các tổ chức tài chính mà còn giúp duy trì sự ổn định của nền kinh tế đang trong thời kì khủng hoảng tài chính.

Ngoài ra, dự đoán khả năng trả nợ của khách hàng bằng học máy là một đề tài vừa mang tính ứng dụng cao trong thực tế, vừa mở ra nhiều cơ hội nghiên cứu và phát triển, đồng thời góp phần thúc đẩy ứng dụng công nghệ trong nhiều lĩnh vực của nghành tài chính, củng cố nền tảng vững để tiếp tục phát triển và mở rộng các dự án mới trong tương lai.

### 2.2 Phương pháp và mục tiêu của đề tài

Phương pháp chính được sử dụng trong đề tài này là xây dựng một ứng dụng hoặc một trang Web dự đoán khả năng trả nợ của khách hàng là có hay không bằng cách sử dụng các mô hình dự đoán để đưa ra quyết định chính xác nhất. Từ đó sẽ có quyết định trong tương lai là có cho vay nữa hay không.

Mục tiêu chính:

* Nghiên cứu, đi sâu và xây dựng nền tảng về mô hình dự đoán – LOGISTIC REGRESSION, XGBOOST.
* Tiền xử lý thành công dữ liệu chứa những thông tin, đặc điểm (xử lý dữ liệu thiếu, loại bỏ giá trị NULL,…).
* Xây dựng thành công mô hình, tinh chỉnh xử lý dự đoán khả năng trả nợ.
* Kiểm thử thực nghiệm thành công với kết quả chính xác cao.

Mục tiêu phụ:

* Xây dựng giao diện ứng dụng hoặc trang Web sử dụng cho người dùng.

### 2.3 Công nghệ áp dụng

* Mô hình hồi quy Logistic: Logistic Regression
* Mô hình XGBoost: eXtreme Gradient Boosting
* Ngôn ngữ Python ( 3.12 64-bit).
* Platform code: Visual Studio Code
* Thư viện xây dựng mô hình: Pandas, Sklearn, LogisticRegression, XGBoost
* Các thư viện khác: matplotlib, numpy, pandas, sklearn, seaborn, joblib, SMOTE,….

- Cấu hình máy:

* Window 11 64-bit.
* System type 64-bit operating system, x64-based processor
* Processor: AMD Ryzen 5 5500U with Radeon Graphics 2.10 GHz.
* Ram: 16GB.

## Cơ sở lý thuyết

### 3.1 Flask là gì?

Flask là loại framework web phổ biến được viết bằng trình lập ngôn ngữ Python. Công nghệ thường được sử dụng để xây dựng trang web từ những ứng dụng đơn giản đến những hệ thống phức tạp hơn.

Flask được thiết kế để hoạt động và mở rộng một cách, đồng thời nó cũng cung cấp các công cụ và thư viện cần thiết để phát triển ứng dụng web hiệu quả. Flask cũng có cộng đồng sáng tạo và hỗ trợ mạnh mẽ từ cộng đồng Python.

Flask được tạo ra bởi Armin Ronacher, một nhà phát triển phần mềm người Đức vào năm 2010. Anh ấy đã phát triển Flask với mục tiêu tạo ra một framework đơn giản nhưng mạnh mẽ để phát triển ứng dụng web bằng Python. Quá trình phát triển Flask được khai thác để đơn giản hóa các yêu cầu và nguồn mã yêu cầu, tạo ra một khung hoạt động và dễ dàng tiếp cận.



Hình 2: Armin Ronacher

Flask được phát triển theo ý tưởng "Micro - Framework", công nghệ tập trung vào các tính năng cơ bản và linh hoạt để người phát triển có thể tùy chỉnh theo nhu cầu cụ thể của họ. Điều này đã giúp Flask trở thành một trong những framework web Python phổ biến nhất và được ưa chuộng trong cộng đồng phát triển phần mềm.

Quá trình phát triển Flask tiếp tục được Armin Ronacher và cộng đồng phát triển mở rộng, nâng cao tính năng và hiệu suất của nó. Flask đã trở thành một công cụ quan trọng cho nhà phát triển Python trong việc xây dựng ứng dụng web hiệu quả và mạnh mẽ.

### 3.2 Tổng hợp các tính năng của Flask Framework

Flask Framework sở hữu một số tính năng quan trọng mà nhà phát triển thường sử dụng để xây dựng hiệu ứng web. Dưới đây là một số tính năng chính của Flask:

Hình 3: Các tính năng của Flask Framework

Nhẹ và dễ sử dụng: Công nghệ có cấu trúc nhẹ nhàng và mã nguồn dễ đọc, giúp người phát triển dễ dàng tiếp cận và tùy chỉnh theo nhu cầu cụ thể của họ.

Định tuyến linh hoạt: Flask cung cấp cơ chế hoạt động định tuyến, cho phép người phát triển xác định các mẫu URL và phân bổ chúng cho các hàm xử lý tương ứng. Điều này giúp quản lý và xử lý yêu cầu HTTP một cách hiệu quả.

Công cụ mẫu: Flask tích hợp Jinja2, đây là một loại trình biên dịch mẫu mạnh mẽ cho phép tạo ra các giao diện người dùng.

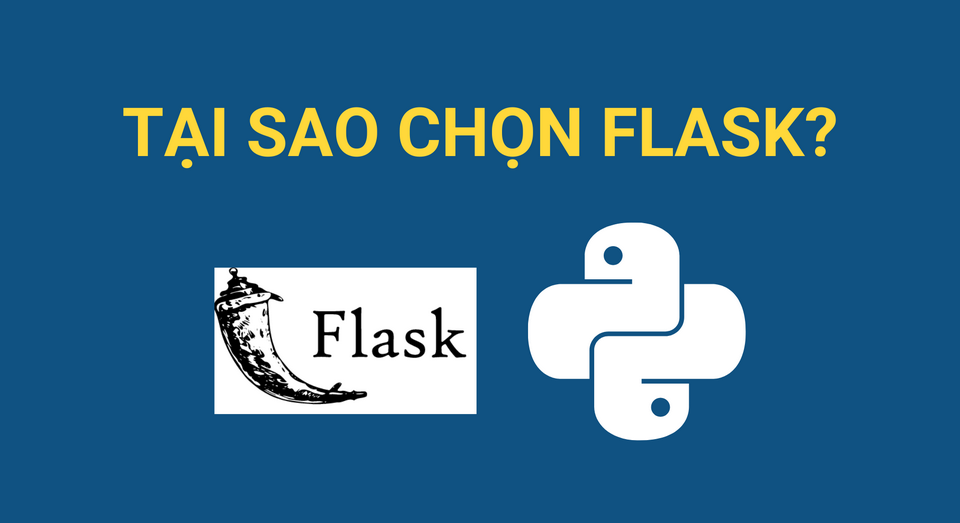
Được mở rộng rộng rãi: Mặc dù mang đặc điểm rút gọn nhưng Flask vẫn có khả năng mở rộng mạnh mẽ thông qua việc sử dụng các tiện ích và thư viện của cộng đồng. Người dùng có thể phân tích các tính năng như xác thực, đăng nhập, điều hướng, cơ sở dữ liệu tương tác và nhiều tính năng khác.• Máy chủ phát triển tích hợp: Flask cung cấp máy chủ phát triển hợp đồng, giúp người phát triển dễ dàng kiểm tra và phát triển ứng dụng mà không cần cấu hình bổ sung.

Gửi yêu cầu RESTful: Flask hỗ trợ xây dựng API và các ứng dụng RESTful theo cách hoạt động và hiệu quả.

Cộng đồng lớn và tích cực: Số lượng người dùng Flask rất đông và luôn nhận được hỗ trợ mạnh mẽ từ cộng đồng Python, điều này giúp người phát triển tìm kiếm thông tin và tài liệu một cách dễ dàng.

### 3.3 Vai trò quan trọng của việc sử dụng Flask

Việc sử dụng Flask mang lại nhiều lợi ích quan trọng đối với nhà phát triển. Đầu tiên, Flask cung cấp một framework linh hoạt và dễ tiếp cận, giúp nhà phát triển tập trung vào logic kinh doanh của ứng dụng mà không bị ràng buộc bởi các quy định cứng nhắc.



Hình 4: Tại sao chọn flask

Flask cũng tích hợp Template Engine mạnh mẽ, Jinja2, để tạo ra giao diện người dùng động một cách dễ dàng. Việc hỗ trợ RESTful Dispatching cũng là một lợi thế lớn khi xây dựng các ứng dụng tương tác và API.

Ngoài ra, Flask có cộng đồng lớn và tích cực nhằm cung cấp nguồn tài liệu phong phú liên quan đến lĩnh vực này. Sự phát triển và mở rộng của Flask thông qua các extension, thư viện cộng đồng sẽ giúp nhà phát triển tiết kiệm thời gian, công sức trong quá trình phát triển ứng dụng web.

### 3.4 Mô hình Logistic Regression

Mô hình hồi quy Logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tổ đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

**Ví dụ:** giả sử bạn muốn đoán xem khách truy cập trang web của bạn sẽ nhấp vào nút thanh toán trong giỏ hàng của bạn hay không. Phân tích hồi quy Logistic xem xét hành vi của khách truy cập trước đây, chẳng hạn như thời gian dành cho trang web và số lượng các mặt hàng trong giỏ hàng. Quá trình phân tích này xác định rằng trước đây nếu khách truy cập dành hơn cụ thể là năm phút trên trang web và thêm hơn ba mặt hàng vào giỏ hàng, họ sẽ nhấp vào nút thành toán. Nhờ vào thông tin này, sau đó hàm hồi quy Logistic có thể dự đoán hành vi của một khách hàng mới truy cập trang web.

Hồi quy Logistic là một kĩ thuật quan trong trong lĩnh vực Trí tuệ nhân tạo và học máy ( AL / ML ). Mô hình hình ML là các chương trình phần mềm có thể được đào tạo để thực hiện các tác vụ xử lý dữ liệu phức tạp mà không cần can thiệp của con người. Mô hình ML được xây dựng bằng hồi quy Logistic có thể giúp các tổ chức thu được thông tin chuyên sâu hữu ích từ dữ liệu kinh doanh của mình. Họ có thể sử dụng những thông tin chuyên sâu này để phân tích dự đoán nhằm giảm chi phí hoạt động, tăng độ hiệu quả và đối chỉnh quy mô nahnh hơn. **Ví dụ:** doanh nghiệp có thể khám phá các mẫu hình cải thiện khả năng giữ chân nhân viên hoặc tạo ra thiết kế sản phẩm mang về nhiều lợi nhuận hơn.

**Tính đơn giản:** Các mo hình hồi quy Logistic ít phức tạp về mặt toán học hơn các phương pháp ML khác. Do đó, bạn có thể triển khai chúng ngay cả khi đội ngũ của bạn không ai có chuyên môn sâu về ML.

**Tốc độ:** Các mô hình hồi quy Logistic có thể xử lý khối lượng lớn dữ liệu ở tốc độ cao bởi chúng cần ít khả năng điện toán hơn, chẳng hạn như bộ nhớ và sức mạnh xử lý. Điều này khiến các mô hình hồi Logistic trở nên lý tưởng đối với những tổ chức đang bắt đầu với các dự án ML để đạt được một số thành tựu nhanh chóng

**Sự linh hoạt:** Bạn có thể sử dụng hồi quy logistic để tìm đáp án cho các câu hỏi có hai hoặc nhiều kết quả hữu hạn. Bạn cũng có thể sử dụng phương pháp này để xử lý trước dữ liệu. Ví dụ: bạn có thể sắp xếp dữ liệu với một phạm vi giá trị lớn, chẳng hạn như giao dịch ngân hàng, thành một phạm vi giá trị hữu hạn, nhỏ hơn nhờ hồi quy logistic. Sau đó, bạn có thể xử lý tập dữ liệu nhỏ hơn này với các kỹ thuật ML khác để phân tích chính xác hơn.

**Khả năng hiển thị:** Phân tích hồi quy logistic cung cấp cho nhà phát triển khả năng nhìn nhận các quy trình phần mềm nội bộ rõ hơn so với các kỹ thuật phân tích dữ liệu khác. Khắc phục sự cố và sửa lỗi cũng trở nên dễ dàng hơn do các phép toán ít phức tạp hơn.

### 3.4 Hồi quy Logistic có những ứng dụng nào ?

Hồi quy logistic có một số ứng dụng thực tế trong nhiều ngành công nghiệp khác nhau.

**Sản xuất:** Các công ty sản xuất áp dụng phân tích hồi quy logistic để ước tính xác suất xảy ra sự cố ở bộ phận trong máy móc. Sau đó, họ sẽ lên lịch bảo trì dựa trên xác suất đã ước tính này để giảm thiểu sự cố trong tương lai.

**Chăm sóc sức khỏe:** Các nhà nghiên cứu y khoa lên kế hoạch điều trị và chăm sóc dự phòng bằng cách dự đoán khả năng mắc bệnh ở bệnh nhân. Họ sử dụng các mô hình hồi quy logistic để so sánh tác động của tiền sử gia đình hoặc của bộ gen lên bệnh tật.

**Tài chính:** Các công ty tài chính phải phân tích các giao dịch tài chính để đề phòng gian lận, xem xét các đơn xin vay và đơn bảo hiểm để đề phòng rủi ro. Những vấn đề này phù hợp với mô hình hồi quy logistic bởi chúng có kết quả cụ thể, chẳng hạn như rủi ro cao hoặc rủi ro thấp và gian lận hoặc không gian lận.

**Bộ phận tiếp thị:** Các công cụ quảng cáo trực tuyến sử dụng mô hình hồi quy logistic để dự đoán xem người dùng sẽ nhấp vào một quảng cáo hay không. Kết quả là, các nhà tiếp thị có thể phân tích phản ứng của người dùng đối với những từ ngữ và hình ảnh khác nhau, tạo ra các quảng cáo hiệu suất cao có khả năng thu hút khách hàng.

### 3.5 Phân tích hồi quy hoạt động như thế nào ?

Hồi quy logistic là một trong số ít các kỹ thuật phân tích hồi quy khác nhau thường được các nhà khoa học dữ liệu sử dụng trong máy học (ML). Để hiểu rõ về hồi quy logistic, trước tiên chúng ta phải hiểu kỹ thuật phân tích hồi quy cơ bản. Dưới đây là một ví dụ về phân tích hồi quy tuyến tính để cho thấy cách thức hoạt động của phân tích hồi quy.

**Xác định câu hỏi:** Bất kỳ quá trình phân tích dữ liệu nào cũng bắt đầu bằng một câu hỏi kinh doanh. Đối với hồi quy logistic, bạn nên giới hạn phạm vi câu hỏi để có được kết quả cụ thể:

Những ngày mưa có ảnh hưởng đến doanh số hàng tháng của chúng ta không? (có hoặc không)

Khách hàng đang thực hiện loại hoạt động thẻ tín dụng nào? (ủy quyền, gian lận hoặc có khả năng gian lận)

**Thu thập dữ liệu lịch sử:** Sau khi xác định câu hỏi, bạn cần xác định các yếu tố dữ liệu có liên quan. Sau đó, bạn sẽ thu thập dữ liệu trước đây cho tất cả các yếu tố. Ví dụ: để trả lời câu hỏi đầu tiên ở trên, bạn có thể thu thập dữ liệu doanh số hàng tháng và số ngày mưa mỗi tháng trong ba năm qua.

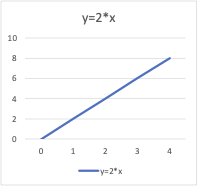
**Đào tạo mô hình phân tích hồi quy:** Bạn sẽ xử lý dữ liệu lịch sử bằng phần mềm hồi quy. Phần mềm sẽ xử lý các điểm dữ liệu khác nhau và kết nối chúng theo phương thức toán học bằng cách sử dụng phương trình. Ví dụ: nếu số ngày mưa trong ba tháng là 3, 5 và 8 còn doanh số trong những tháng đó là 8, 12 và 18, thuật toán hồi quy sẽ kết nối các yếu tố này với phương trình - Doanh số = 2\*(Số ngày mưa) + 2

**Dự đoán các giá trị không xác định:** Đối với các giá trị không xác định, phần mềm sẽ áp dụng phương trình để dự đoán. Nếu bạn biết rằng tháng 7 sẽ có sáu ngày mưa, phần mềm sẽ ước tính giá trị doanh số trong tháng 7 là 14.

### 3.6 Mô hình hồi quy Logistic hoạt động như thế nào ?

**Để hiểu rõ về mô hình hồi quy logistic, trước tiên chúng ta phải hiểu các phương trình và biến.**

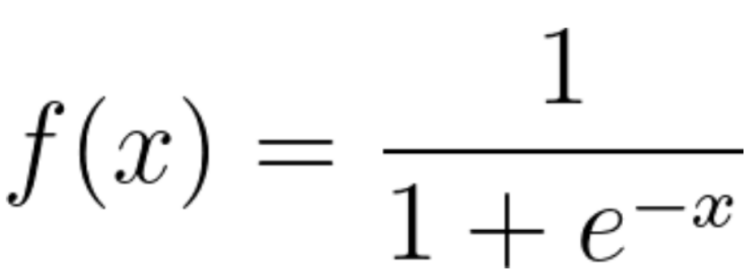
**Phương trình:** Trong toán học, phương trình cho ta mối quan hệ giữa hai biến: x và y. Bạn có thể sử dụng các phương trình hoặc hàm này để vẽ đồ thị theo trục x và trục y bằng cách nhập các giá trị khác nhau của x và y. Ví dụ: nếu bạn vẽ đồ thị cho hàm y = 2\*x, bạn sẽ có một đường thẳng như hình dưới đây. Do đó hàm này còn được gọi là hàm tuyến tính.



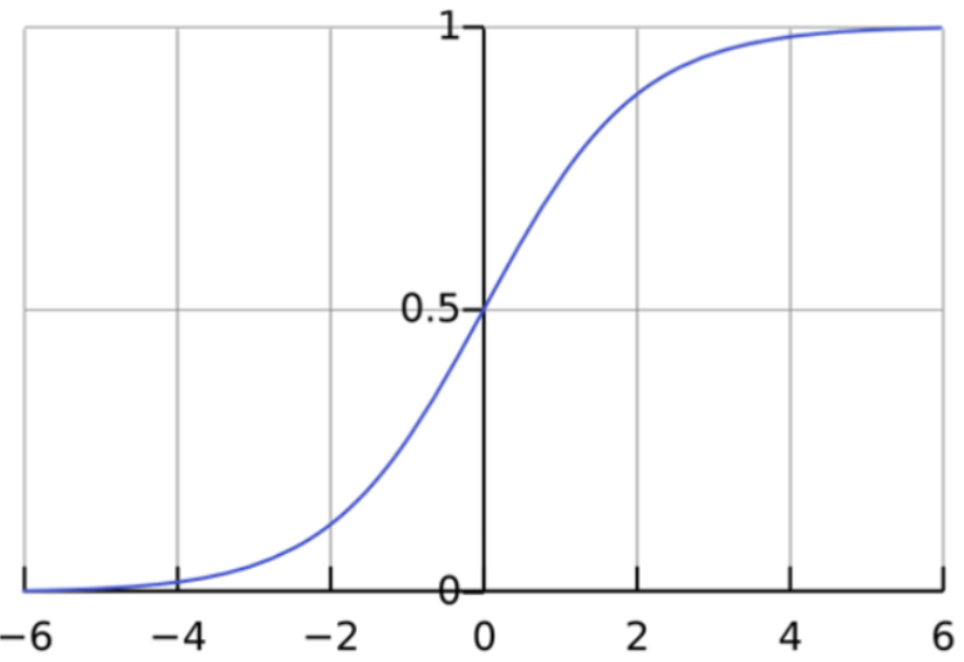
Hình 5: Phương trình quan hệ

**Biến:** Trong thống kê, biến là các yếu tố dữ liệu hoặc thuộc tính có giá trị khác nhau. Bất kỳ phân tích nào cũng có một số biến nhất định là biến độc lập hoặc biến giải thích. Những thuộc tính này là nguyên nhân của một kết quả. Các biến khác là biến phụ thuộc hoặc biến đáp ứng; giá trị của chúng phụ thuộc vào các biến độc lập. Nhìn chung, hồi quy logistic khám phá cách các biến độc lập ảnh hưởng đến một biến phụ thuộc bằng cách xem xét các giá trị dữ liệu lịch sử của cả hai biến.

**Hàm hồi quy Logistic:** Hồi quy logistic là một mô hình thống kê sử dụng hàm logistic, hay hàm logit trong toán học làm phương trình giữa x và y. Hàm logit ánh xạ y làm hàm sigmoid của x.



\* Nếu vẽ phương trình hồi quy logistic này, bạn sẽ có một đường cong hình chữ S như hình dưới đây.



Hình 6: Phương trình hồi quy

Như bạn có thể thấy, hàm logit chỉ trả về các giá trị giữa 0 và 1 cho biến phụ thuộc, dù giá trị của biến độc lập là gì. Đây là cách hồi quy logistic ước tính giá trị của biến phụ thuộc. Phương pháp hồi quy logistic cũng lập mô hình phương trình giữa nhiều biến độc lập và một biến phụ thuộc.

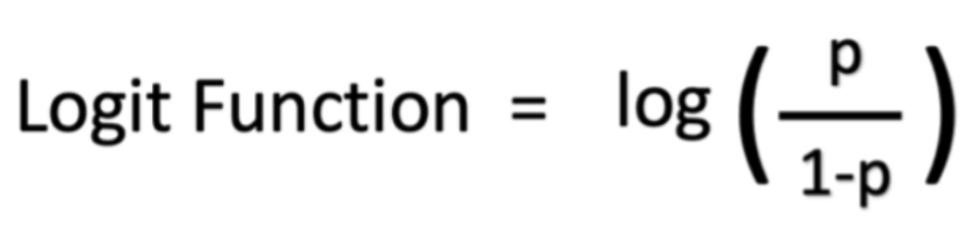
**Phân tích hồi quy logistic với nhiều biến độc lập:** Trong nhiều trường hợp, nhiều biến giải thích ảnh hưởng đến giá trị của biến phụ thuộc. Để lập mô hình các tập dữ liệu đầu vào như vậy, công thức hồi quy logistic phải giả định mối quan hệ tuyến tính giữa các biến độc lập khác nhau. Bạn có thể sửa đổi hàm sigmoid và tính toán biến đầu ra cuối cùng như sau:



Ký hiệu β đại diện cho hệ số hồi quy. Mô hình logit có thể đảo ngược tính toán các giá trị hệ số này khi bạn cho nó một tập dữ liệu thực nghiệm đủ lớn có các giá trị đã xác định của cả hai biến phụ thuộc và biến độc lập.

**Log của tỷ số odds:** Mô hình logit cũng có thể xác định tỷ số thành công trên thất bại hay log của tỷ số odds. Ví dụ: nếu bạn đang chơi poker với bạn bè và thắng bốn ván trên mười ván, tỷ số chiến thắng của bạn là bốn phần sáu, hoặc 4/6, và đó là tỷ số thành công trên thất bại của bạn. Mặt khác, xác suất thắng là 4/10.

Về mặt toán học, tỷ số odds về mặt xác suất của bạn là p/(1 - p) và log của tỷ số odds là log (p/(1 - p)). Bạn có thể biểu diễn hàm logistic bằng log của tỷ số odds như hình dưới đây:



### 3.7 Phân tích hồi quy logistic có những loại nào?

Có ba cách tiếp cận phân tích hồi quy logistic dựa trên kết quả của biến phụ thuộc.

**Hồi quy logistic nhị phân:** Hồi quy logistic nhị phân phù hợp với các vấn đề phân lớp nhị phân chỉ có hai kết quả có thể xảy ra. Biến phụ thuộc chỉ có thể có hai giá trị, chẳng hạn như có và không hoặc 0 và 1. Dù hàm logistic tính toán một phạm vi giá trị giữa 0 và 1, mô hình hồi quy nhị phân vẫn sẽ làm tròn kết quả đến các giá trị gần nhất. Nói chung, kết quả dưới 0,5 sẽ được làm tròn thành 0 và kết quả trên 0,5 sẽ được làm tròn thành 1, do đó hàm logistic trả về một kết quả nhị phân.

**Hồi quy logistic đa thức:** Hồi quy đa thức có thể phân tích các vấn đề có một số kết quả có thể xảy ra, miễn là số kết quả hữu hạn. Ví dụ: kỹ thuật này có thể dự đoán xem giá nhà sẽ tăng 25%, 50%, 75% hay 100% dựa trên dữ liệu dân số, nhưng sẽ không thể dự đoán được giá trị chính xác của một ngôi nhà. Hồi quy logistic đa thức hoạt động bằng cách ánh xạ các giá trị kết quả cho các giá trị khác nhau giữa 0 và 1. Hàm logistic có thể trả về một khoảng dữ liệu liên tục như 0,1, 0,11, 0,12, v.v., do đó hồi quy đa thức cũng nhóm đầu ra đến các giá trị gần nhất có thể có.

**Hồi quy logistic thứ tự:** Hồi quy logistic thứ tự, hay mô hình logit có thứ tự, là một loại hồi quy đa thức đặc biệt cho các vấn đề trong đó các số đại diện cho các bậc chứ không phải là giá trị thực tế. Ví dụ: bạn sẽ sử dụng hồi quy thứ tự để dự đoán đáp án cho câu hỏi khảo sát yêu cầu khách hàng đánh giá dịch vụ của bạn ở mức kém, ổn, tốt hoặc xuất sắc dựa trên một giá trị số, chẳng hạn như số lượng mặt hàng họ mua từ bạn trong năm.

**3.6 Làm thế nào để so sánh hồi quy Logistic với các kỹ thuật ML khác ?**

Hai kỹ thuật phân tích dữ liệu phổ biến là phân tích hồi quy tuyến tính và học sâu.

**Phân tích hồi quy tuyến tính:** Như đã giải thích ở trên, hồi quy tuyến tính lập mô hình mối quan hệ giữa các biến phụ thuộc và độc lập bằng tổ hợp tuyến tính. Phương trình hồi quy tuyến tính là:



**Hồi quy Logistic so với hồi quy tuyến tính:** Hồi quy tuyến tính dự đoán một biến phụ thuộc liên tục bằng một tập hợp các biến độc lập cho trước. Một biến liên tục có thể có một phạm vi giá trị, chẳng hạn như giá cả hoặc độ tuổi. Do đó, hồi quy tuyến tính có thể dự đoán giá trị thực của biến phụ thuộc. Kỹ thuật này có thể trả lời các câu hỏi như "Giá gạo sau 10 năm nữa sẽ là bao nhiêu?" Không giống như hồi quy tuyến tính, hồi quy logistic là một thuật toán phân loại. Kỹ thuật này không thể dự đoán giá trị thực sự cho dữ liệu liên tục. Kỹ thuật này có thể trả lời các câu hỏi như "Liệu giá gạo trong 10 năm nữa có tăng 50% hay không?"

**Học sâu:** Học sâu sử dụng mạng nơ-ron hoặc các thành phần phần mềm mô phỏng bộ não con người để phân tích thông tin. Các phép toán học sâu dựa trên khái niệm toán học của vectơ.

**Hồi quy Logistic so với học sâu:** Hồi quy logistic ít phức tạp và có cường độ điện toán ít hơn so với học sâu. Quan trọng hơn là nhà phát triển không thể điều tra hoặc sửa đổi các phép toán học sâu do tính chất phức tạp và dựa trên máy móc của chúng. Mặt khác, các phép toán hồi quy logistic lại minh bạch và dễ khắc phục sự cố hơn.

**3.7 Mô hình XGBoost**

XGBoost, hay eXtreme Gradient Boosting, là một thuật toán XGBoost trong thuật toán học máy theo học tập tổng hợp. Thuật toán này rất phổ biến đối với các tác vụ học có giám sát, chẳng hạn như hồi quy và phân loại. XGBoost xây dựng một mô hình dự đoán bằng cách kết hợp các dự đoán của nhiều mô hình riêng lẻ, thường là cây quyết định, theo cách lặp lại.

Thuật toán hoạt động bằng cách tuần tự thêm các học viên yếu vào nhóm, với mỗi học viên mới tập trung vào việc sửa lỗi do các học viên hiện tại mắc phải. Thuật toán sử dụng kỹ thuật tối ưu hóa giảm dần độ dốc để giảm thiểu hàm mất mát được xác định trước trong quá trình đào tạo.

Các tính năng chính của thuật toán XGBoost bao gồm khả năng xử lý các mối quan hệ phức tạp trong dữ liệu, các kỹ thuật chính quy hóa để ngăn ngừa tình trạng quá khớp và kết hợp xử lý song song để tính toán hiệu quả.

\* **Tại sao nên học theo nhóm:** XGBoost là một phương pháp học tập tổng hợp. Đôi khi, việc dựa vào kết quả của chỉ một mô hình học máy có thể không đủ. Học tập tổng hợp cung cấp một giải pháp có hệ thống để kết hợp sức mạnh dự đoán của nhiều người học. Kết quả là một mô hình duy nhất cung cấp đầu ra tổng hợp từ một số mô hình.

Các mô hình tạo nên tập hợp, còn được gọi là người học cơ sở, có thể đến từ cùng một thuật toán học hoặc các thuật toán học khác nhau. Bagging và increasing đóng vai trò là hai người học tập tập hợp được sử dụng rộng rãi. Mặc dù bạn có thể áp dụng các kỹ thuật này với một số mô hình thống kê, nhưng cây quyết định chiếm ưu thế trong việc sử dụng chúng.

Chúng ta hãy thảo luận ngắn gọn về bagging trước khi đi sâu hơn vào khái niệm tăng cường gradient.

**\* Đóng bao:** Trong khi cây quyết định là một trong những mô hình dễ diễn giải nhất, chúng thể hiện hành vi rất biến đổi. Hãy xem xét một tập dữ liệu đào tạo duy nhất mà chúng ta chia ngẫu nhiên thành hai phần. Bây giờ, hãy sử dụng từng phần để đào tạo một cây quyết định nhằm thu được hai mô hình.

Khi chúng ta lắp cả hai mô hình này, chúng sẽ tạo ra các kết quả khác nhau. Cây quyết định thể hiện độ biến thiên cao do hành vi này. Tổng hợp đóng gói hoặc tăng cường giúp giảm độ biến thiên ở bất kỳ người học nào. Một số cây quyết định được tạo song song tạo thành người học cơ sở của kỹ thuật đóng gói. Dữ liệu được lấy mẫu với sự thay thế được đưa vào những người học này để đào tạo. Dự đoán cuối cùng là đầu ra trung bình từ tất cả người học.

**\* Tăng cường:** Trong quá trình tăng cường, các cây được xây dựng theo trình tự để mỗi cây tiếp theo có mục đích giảm lỗi của cây trước đó. Mỗi cây học từ cây tiền nhiệm và cập nhật các lỗi còn lại. Do đó, cây phát triển tiếp theo trong trình tự sẽ học từ phiên bản cập nhật của các phần còn lại.

Những người học cơ bản trong việc thúc đẩy là những người học yếu trong đó độ lệch cao và sức mạnh dự đoán chỉ tốt hơn một chút so với việc đoán ngẫu nhiên. Mỗi người học yếu này đóng góp một số thông tin quan trọng để dự đoán, cho phép kỹ thuật thúc đẩy tạo ra một người học mạnh bằng cách kết hợp hiệu quả những người học yếu này. Người học mạnh cuối cùng làm giảm cả độ lệch và phương sai.

Ngược lại với các kỹ thuật bagging như Random Forest, increasing sử dụng các cây có ít lần chia tách hơn. Những cây nhỏ như vậy, không quá sâu, có khả năng diễn giải cao. Bạn có thể lựa chọn tối ưu các tham số như số lượng cây hoặc số lần lặp, tốc độ học của gradient increasing và độ sâu của cây thông qua các kỹ thuật xác thực như k-fold cross-validation. Có số lượng cây lớn có thể dẫn đến overfitting. Vì vậy, cần phải lựa chọn cẩn thận các tiêu chí dừng để increasing.

**\* Kỹ thuật tăng cường Gradient:** Kỹ thuật tổng hợp tăng cường độ dốc bao gồm ba bước đơn giản:

• Một mô hình ban đầu F 0 được xác định để dự đoán biến mục tiêu y. Mô hình này sẽ được liên kết với một phần dư (y – F 0 )

• Một mô hình h 1 mới phù hợp với các phần dư từ bước trước

• Bây giờ, F 0 và h 1 được kết hợp để tạo ra F 1 , phiên bản tăng cường của F 0. Sai số bình phương trung bình từ F 1 sẽ thấp hơn sai số từ F 0:



Để cải thiện hiệu suất của F 1 , chúng ta có thể mô hình hóa theo phần dư của F 1 và tạo ra một mô hình mới F 2:



Điều này có thể được thực hiện với 'm' lần lặp, cho đến khi phần dư được giảm thiểu hết mức có thể:



Ở đây, người học cộng tính không làm xáo trộn các chức năng được tạo ra ở các bước trước đó. Thay vào đó, họ truyền đạt thông tin của riêng mình để giảm lỗi.

**\* Cách hoạt động của XGBoost:**

Khởi tạo:

• Bắt đầu với một dự đoán ban đầu (thường là trung bình của nhãn mục tiêu trong bài toán hồi quy hoặc xác suất trong bài toán phân loại).

Tạo cây quyết định:

• Một cây quyết định được xây dựng để giảm thiểu sai số của dự đoán hiện tại.

• Sử dụng hàm mất mát (loss function) để đo lường mức độ sai số.

Cập nhật dự đoán:

• Kết hợp các dự đoán từ cây quyết định mới với các dự đoán trước đó bằng cách sử dụng hệ số learning rate.

Lặp lại:

• Tiếp tục xây dựng cây mới, tập trung vào việc giảm các lỗi còn lại.

Hội tụ:

• Quá trình dừng lại khi không còn cải thiện đáng kể hoặc đạt đến số vòng lặp tối đa.

**\* Các tham số quan trọng:**

**max\_depth**: Độ sâu tối đa của cây quyết định.

**learning\_rate (eta):** Tốc độ học của mô hình, giúp kiểm soát mức độ tác động của từng cây.

**n\_estimators:** Số lượng cây quyết định trong mô hình.

**subsample:** Tỷ lệ mẫu sử dụng cho mỗi cây.

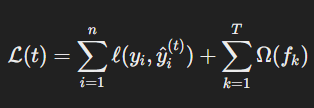
**colsample\_bytree:** Tỷ lệ cột được sử dụng khi xây dựng mỗi cây.

**gamma:** Mức độ tối thiểu của cải thiện mất mát để tạo thêm một nút phân tách trong cây.

**lambda và alpha:** Các tham số regularization L2 và L1.

**\* Hàm mục tiêu trong XGBoost**

Hàm mục tiêu được tối ưu hóa trong XGBoost bao gồm hai phần:



Trong đó:

• T: Số nút trong cây

• λ: Tham số chính quy hóa L2

• wj: Trọng số tại nút lá j

• γ: Độ phức tạp của cây

**\* Cập nhật dự đoán sử dụng Gradient Boosting**

Ở vòng lặp t, mục tiêu của XGBoost là tìm hàm ft(x) để giảm thiểu hàm mất mát. Dự đoán mới được tính như sau:



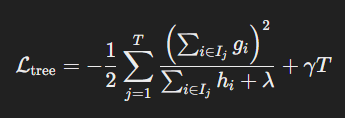
Trong đó:

• Yi(t-1): Dự đoán tại vòng lặp trước đó

• Ft(xi): Giá trị dự đoán của cây t cho mẫu xi

**\* Hàm mục tiêu tại cây quyết định:**

Hàm mục tiêu tổng quát cho một cây quyết định có thể được viết lại như sau:

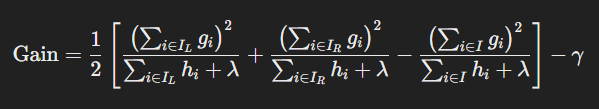


Trong đó:

• γT: Penalization cho số lượng nút lá T trong cây

**\* Split điểm tối ưu:**

XGBoost quyết định chia điểm (split point) tại mỗi nút dựa trên độ tăng lợi ích (Gain):



Trong đó:

• IL: Tập dữ liệu tại nhánh trái sau khi split

• IR: Tập dữ liệu tại nhánh phải sau khi split

• I: Tập dữ liệu tại nút trước khi split

Nếu Gain > 0, việc split sẽ được thực hiện

# III. XÂY DỰNG MÔ HÌNH

## Tổng quan phương pháp xây dựng

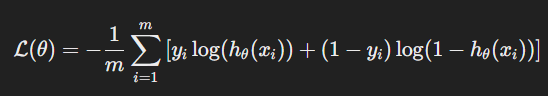
Dưới đây là các bước chính để xây dựng một mô hình Logistic Regression:

#### ****Bước 1: Thu thập và tiền xử lý dữ liệu****

* **Thu thập dữ liệu**: Chuẩn bị tập dữ liệu huấn luyện, đảm bảo có nhãn yyy và các đặc trưng xxx.
* **Xử lý dữ liệu thiếu**: Loại bỏ hoặc thay thế các giá trị bị thiếu.
* **Chuẩn hóa dữ liệu**: Biến đổi các đặc trưng để tránh vấn đề tỷ lệ chênh lệch, ví dụ: chuẩn hóa về khoảng [0,1][0, 1][0,1] hoặc chuẩn hóa Z-score.

**Bước 2: Định nghĩa hàm mất mát**

Sử dụng hàm mất mát log-likelihood để đánh giá mức độ khớp của mô hình:



Trong đó:

m: Số lượng mẫu trong tập huấn luyện

Yi: Nhãn thực tế của mẫu i ( 0 hoặc 1 )

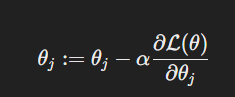
H0(xi): Xác suất dự đoán của mẫu I

Hàm mất mát này tối đa hóa khả năng (likelihood) của tham số θ dựa trên dữ liệu quan sát.

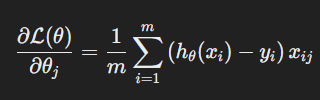
**Bước 3: Tối ưu hóa tham số**

Tìm tham số θ tốt nhất bằng cách tối thiểu hóa hàm mất mát. Phương pháp phổ biến nhất là gradient descent hoặc các biến thể như stochastic gradient descent (SGD).

• Gradient Descent:



• Với:



Trong đó:

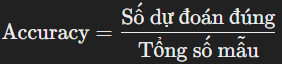
• α: Tốc độ học ( learning rate )

• xij: Giá trị của đặc trưng j trong mẫu I

**Bước 4: Đánh giá mô hình**

Sử dụng các chỉ số để đánh giá hiệu suất mô hình:

• Accuracy ( Độ chính xác ):

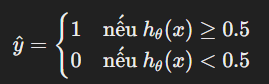


• Precision, Recall, F1-score: Đặc biệt quan trọng trong các bài toán có dữ liệu không cân bằng

• ROC-AUC: Đánh giá khả năng phân loại tốt trên các ngưỡng khác nhau

**Bước 5: Triển khai mô hình**

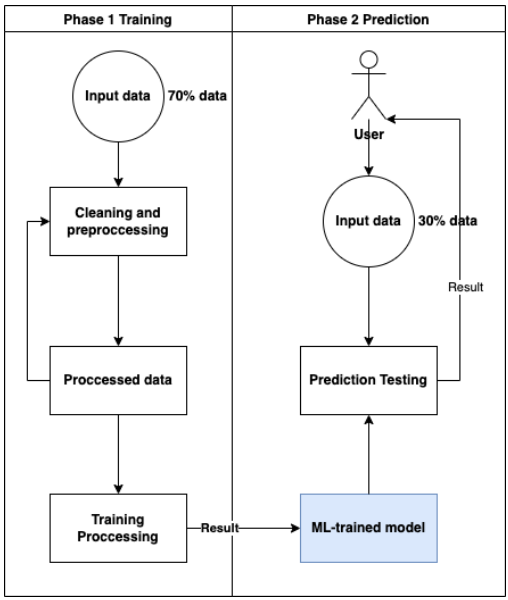
• Sử dụng mô hình để dự đoán trên dữ liệu mới:



• Triển khai mô hình trong môi trường thực tế để phục vụ ứng dụng như phân loại email, phát hiện gian lận, hoặc dự đoán rủi ro.

**Bước 6: Xây dựng giao diện người dùng**

• Xây dựng giao diện người dùng bằng thư viện PyQt5 và Web Flask



Hình 7: Biểu đồ quá trình huấn luyện

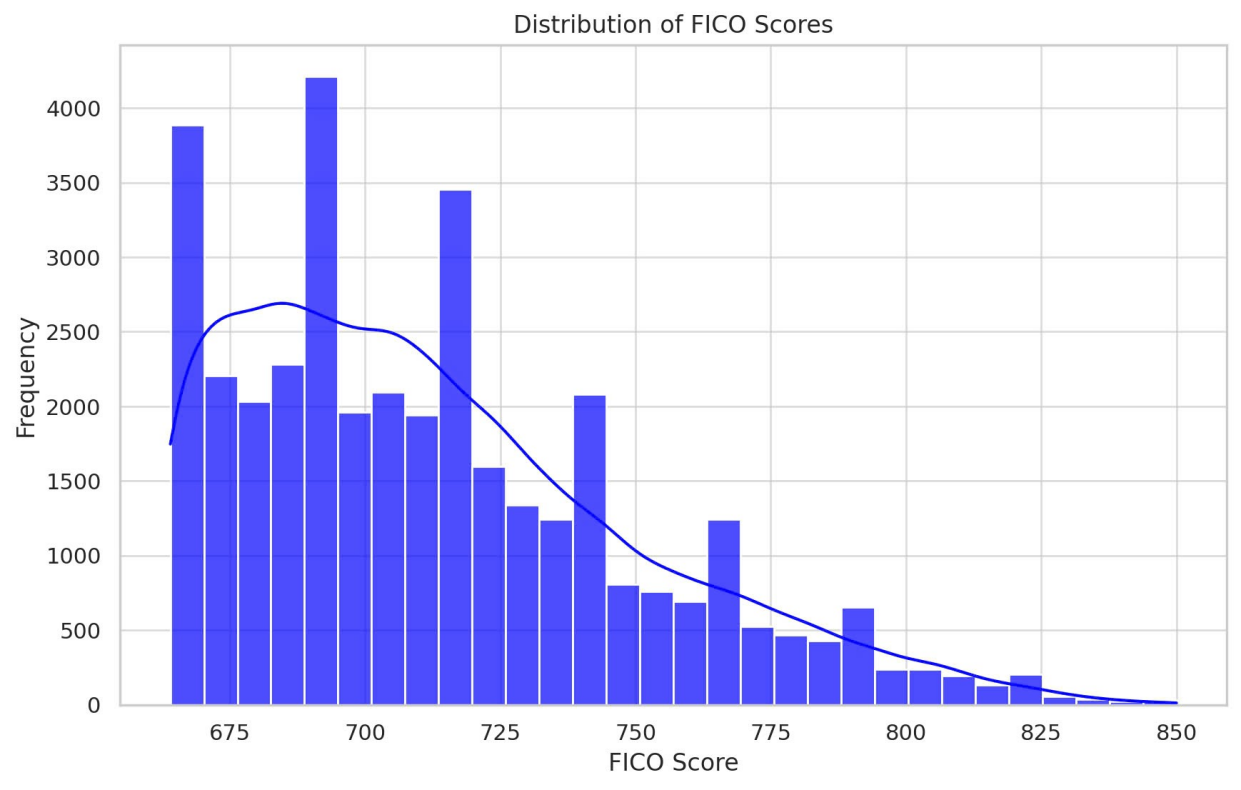
## Quy trình xây dựng mô hình

### 2.1 Thu thập dữ liệu

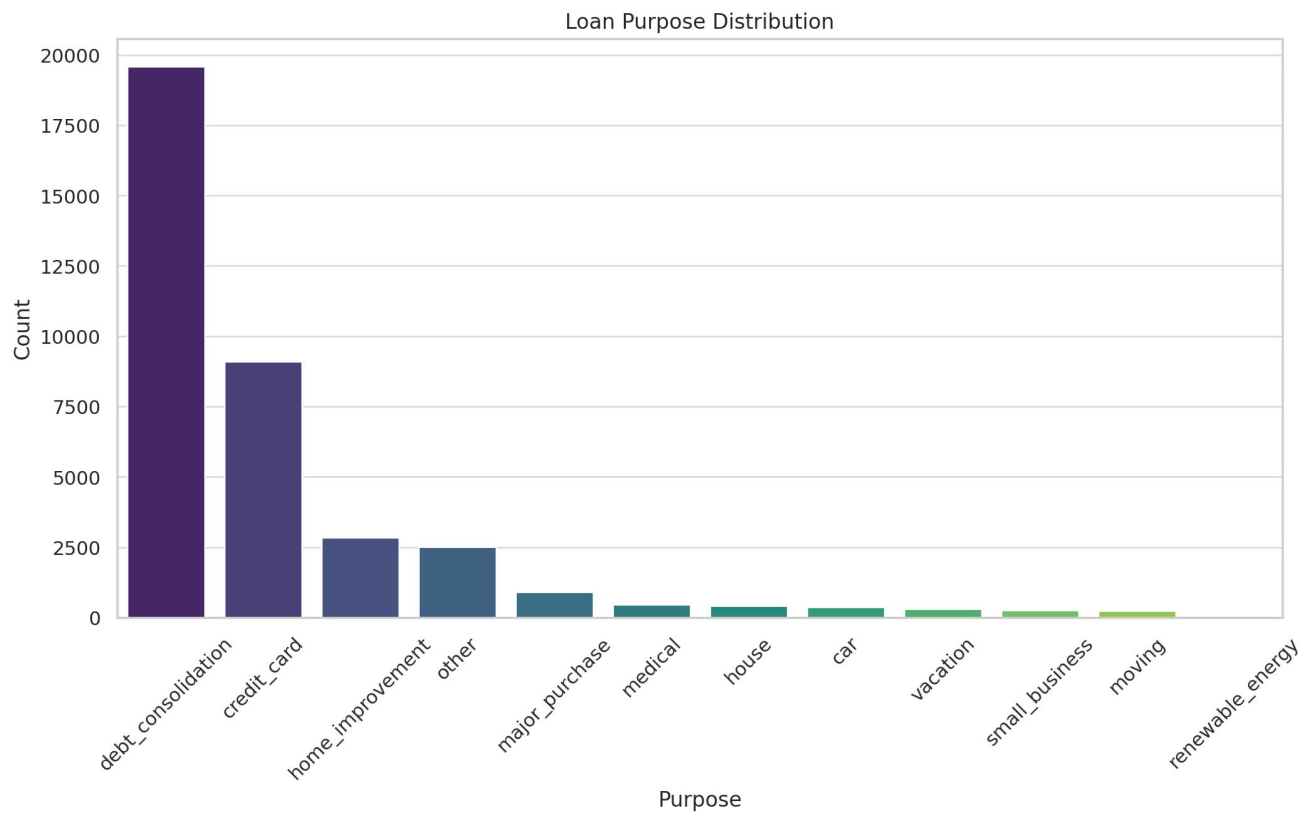
Dữ liệu có 37,066 dòng và 14 cột. Mỗi dòng đại diện cho thông tin vay của một khách hàng, bao gồm các đặc điểm như điểm tín dụng (FICO), tỷ lệ lãi suất, mục đích vay, tỷ lệ nợ trên thu nhập (DTI), và trạng thái khoản vay (trả đầy đủ hay không).

### 2.2 Phân tích sơ bộ và trực quan hóa dữ liệu:

Chi tiết hơn sẽ ở phần trực quan hóa như sau:



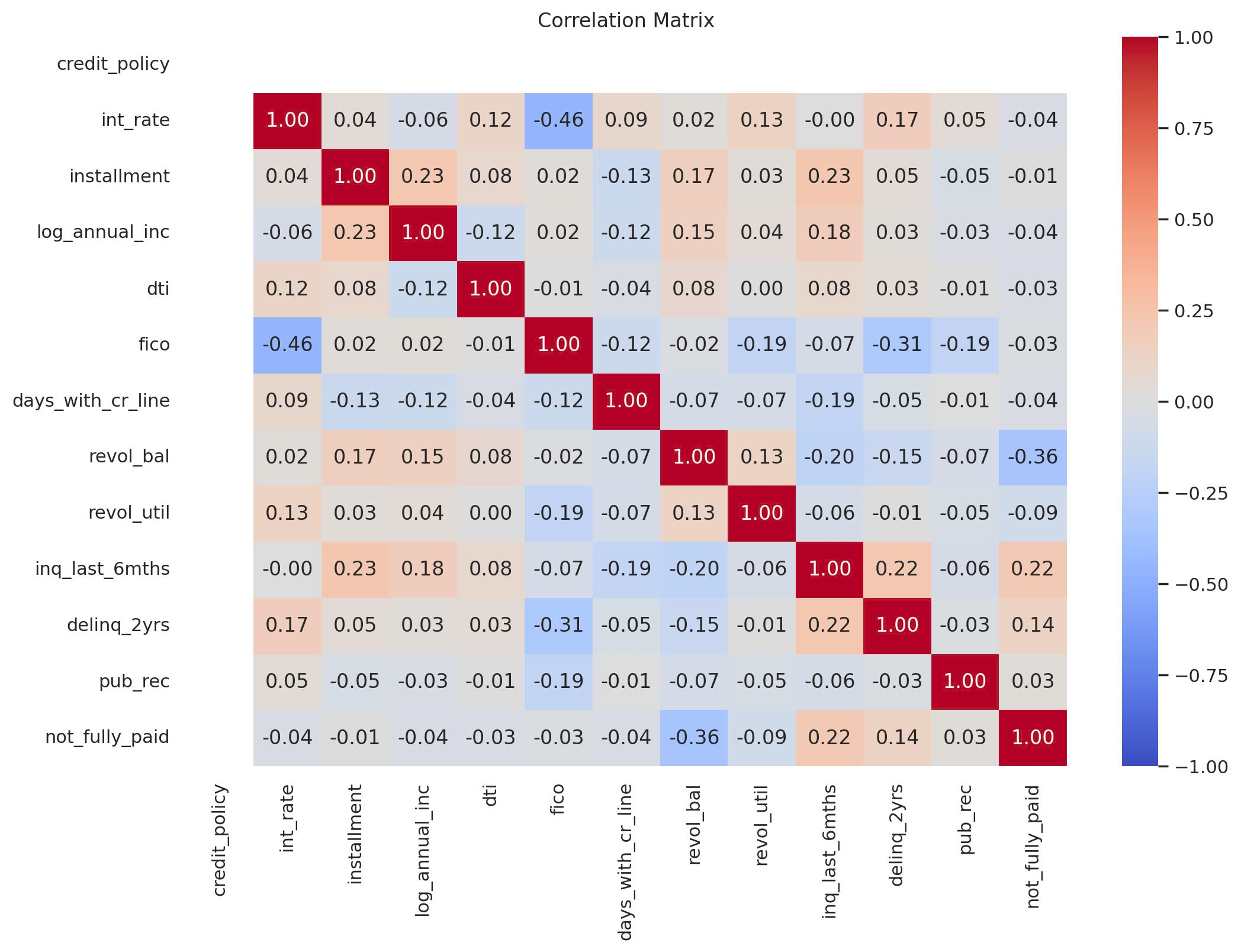
Hình 8: Trực quan hóa phân phối điểm FICO



Hình 9: Trực quan hóa phân phối mục đích vay



Hình 10: Trực quan hóa mối quan hệ giữa điểm FICO và trạng thái trả nợ



Hình 11: Trực quan hóa ma trận tương quan

### 2.3 Tiền xử lý dữ liệu

- **Xử lý giá trị bị thiếu:** Kiểm tra các cột có giá trị thiếu và áp dụng phương pháp xử lý phù hợp:

+ dti: Có 100 giá trị thiếu. Có thể điền bằng giá trị trung bình hoặc trung vị.

+ revol\_util: Có 12 giá trị thiếu. Điền bằng giá trị trung bình.

+ delinq\_2yrs: Có 34 giá trị thiếu. Điền bằng 0 (không có lần chậm thanh toán).

- **Mã hóa biến phân loại:** Cột purpose (mục đích vay) là một biến phân loại (categorical). Cần chuyển đổi thành dạng số:

+ Sử dụng One-Hot Encoding hoặc Label Encoding.

- **Chuẩn hóa dữ liệu:** Để cải thiện hiệu quả mô hình học máy:

+ Chuẩn hóa các biến số học (fico, int\_rate, installment, v.v.) về cùng thang giá trị (0-1 hoặc chuẩn hóa Z-score).

- **Xử lý dữ liệu ngoại lại:**

+ Kiểm tra các cột như dti, int\_rate, fico, v.v. để phát hiện và xử lý giá trị ngoại lai (nếu cần).

- **Tách dữ liệu đầu vào và đầu ra:**

+ Biến mục tiêu là not\_fully\_paid.

+ Tách các cột còn lại làm dữ liệu đầu vào.

# IV. THỰC NGHIỆM

## 1. Xây dựng thực nghiệm

Sau khi huấn luyện cả hai mô hình Logistic Regression và Random Forests trên tập huấn luyện, chúng tôi tiến hành đánh giá hiệu năng của chúng trên tập kiểm tra độc lập. Việc đánh giá này sử dụng nhiều chỉ số khác nhau để có cái nhìn toàn diện về hiệu năng của từng mô hình.

* **Độ chính xác (Accuracy)**

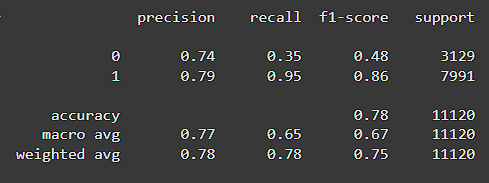
Độ chính xác là tỉ lệ giữa số lượng dự đoán đúng và tổng số dự đoán. Trong thực nghiệm này, Logistic Regression đạt độ chính xác 78%, trong khi Random Forests đạt độ chính xác 85%. Điều này cho thấy Logistic Regression dự đoán đúng cao hơn Random Forests .

* **Độ phủ (Recall)**

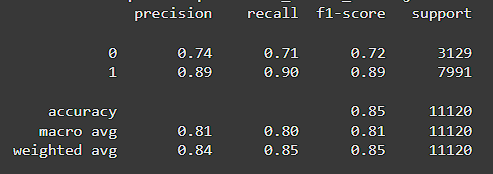
Độ phủ là tỉ lệ giữa số lượng dự đoán dương tính đúng và tổng số mẫu thực sự là dương tính. Chỉ số này cho biết trong số những mẫu thực sự là dương tính, có bao nhiêu mẫu được dự đoán đúng. Logistic Regression đạt độ phủ 0.95, trong khi Random Forests đạt độ phủ 0.90.

* **F1-score**

F1-score là trung bình điều hòa giữa độ chính xác và độ phủ. Chỉ số này cân bằng giữa độ chính xác và độ phủ. Logistic Regression đạt F1-score 0.85, trong khi Random Forests đạt F1-score 0.78



Hình 12: Kết quả huấn luyện model Logistic Regression

**

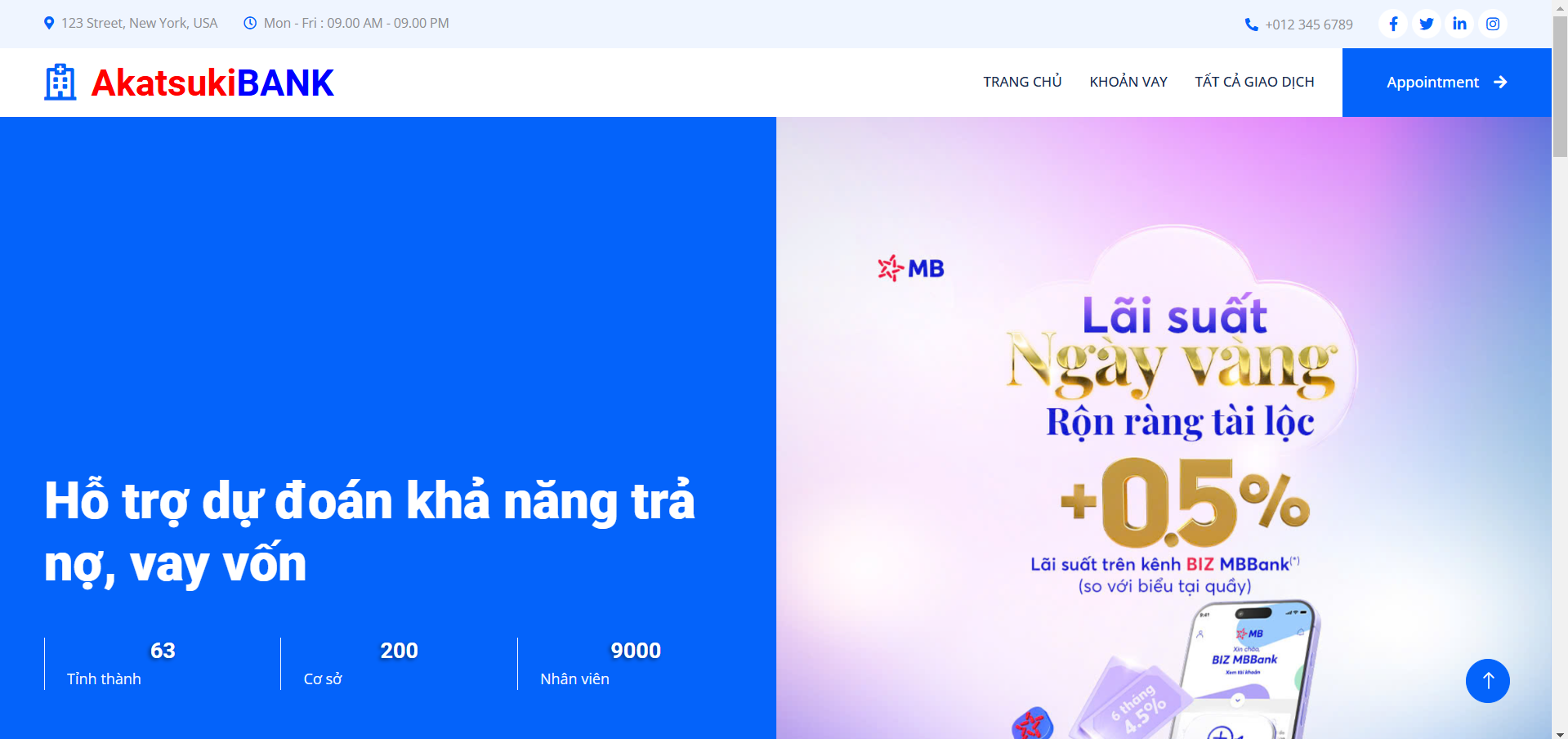
Hình 13: Kết quả huấn luyện model Random Forests

Nhìn chung, Logistic Regression cho kết quả tốt hơn so với Random Forests trên hầu hết các chỉ số đánh giá. Điều này cho thấy rằng, trong trường hợp cụ thể này, Logistic Regression là mô hình phù hợp hơn cho bài toán phân loại nhị phân trên tập dữ liệu được sử dụng. Sự khác biệt về hiệu năng có thể do nhiều yếu tố, bao gồm tính chất của dữ liệu, sự phức tạp của mô hình và việc điều chỉnh tham số. Chúng ta sẽ thảo luận chi tiết hơn về những yếu tố này trong phần thảo luận.

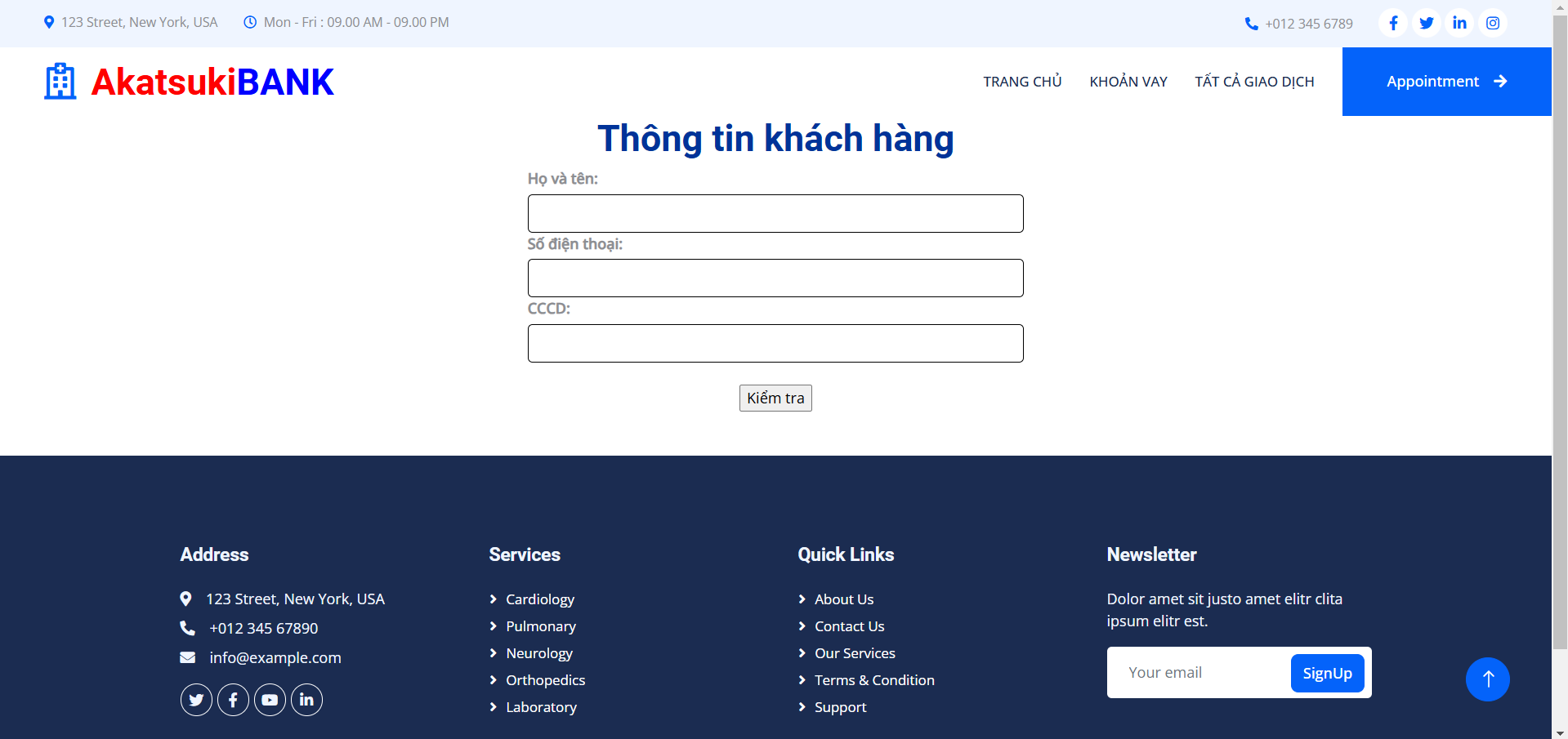
* **Xây dựng giao diện với thư viện flask**

Thư viện này được thiết kế để người sử dụng có thể nhập các thông tin được cung cấp từ bên khách hàng trên Website hoặc của một tổ chức tín dụng để thực hiện dự đoán

Giao diện được cải thiện hoàn chỉnh:

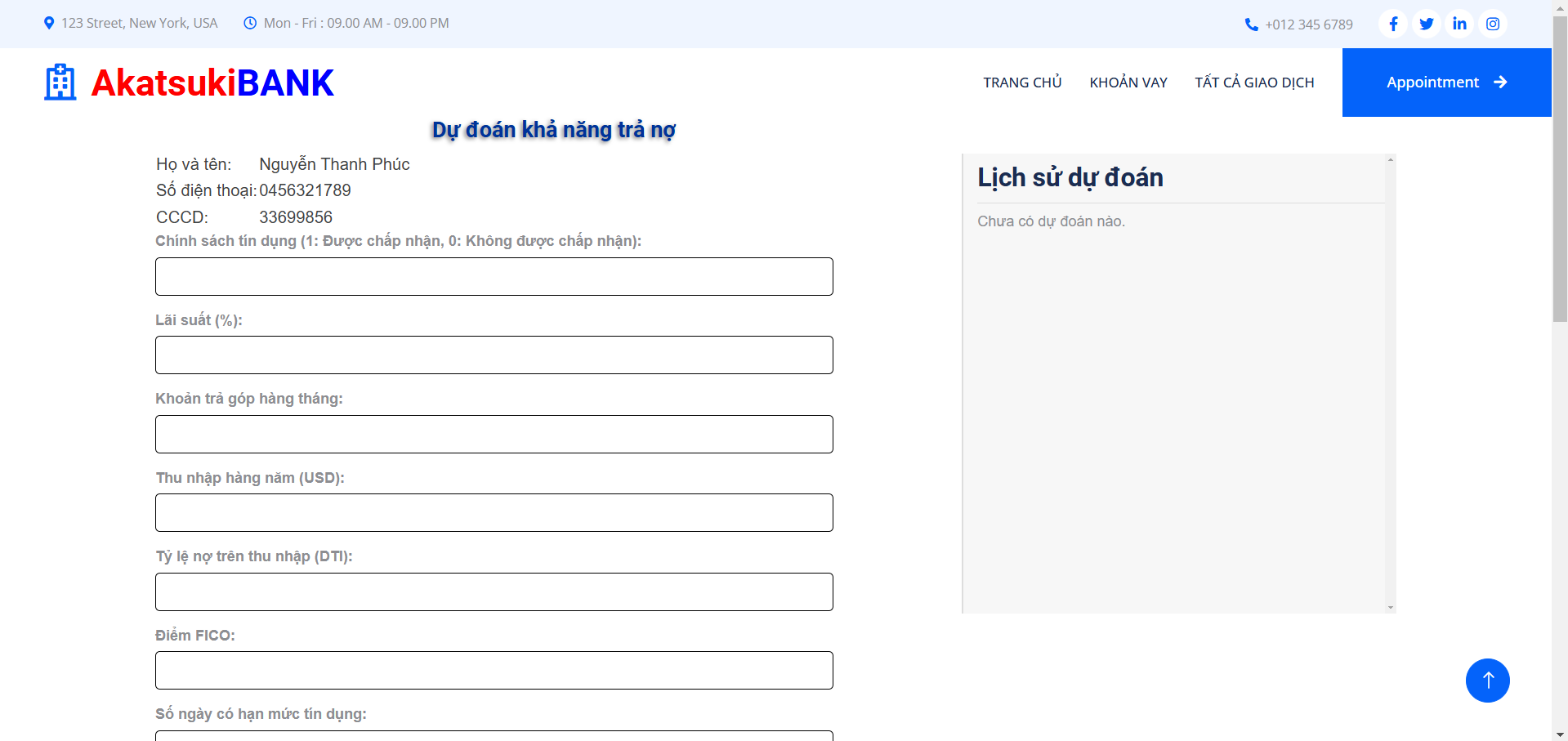


Hình 14: Giao diện cải tiến hoàn chỉnh



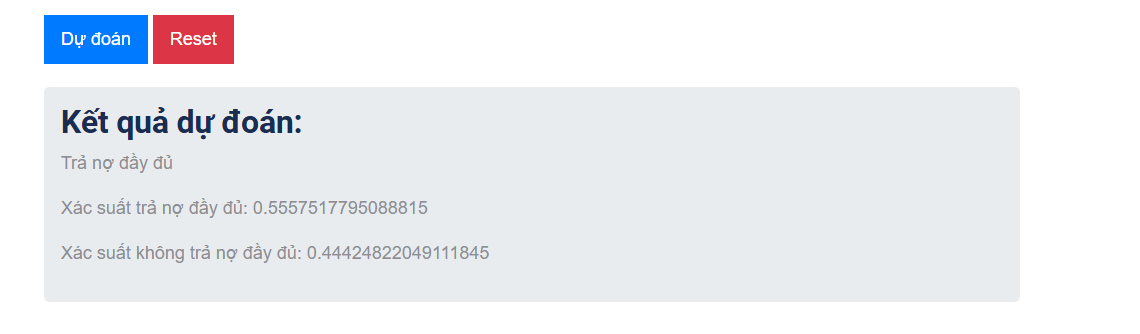
Hình 15: Nhập thông tin

Khi bấm vào mục “ Khoản vay “ sẽ hiện ra giao diện nhập thông tin khách hàng



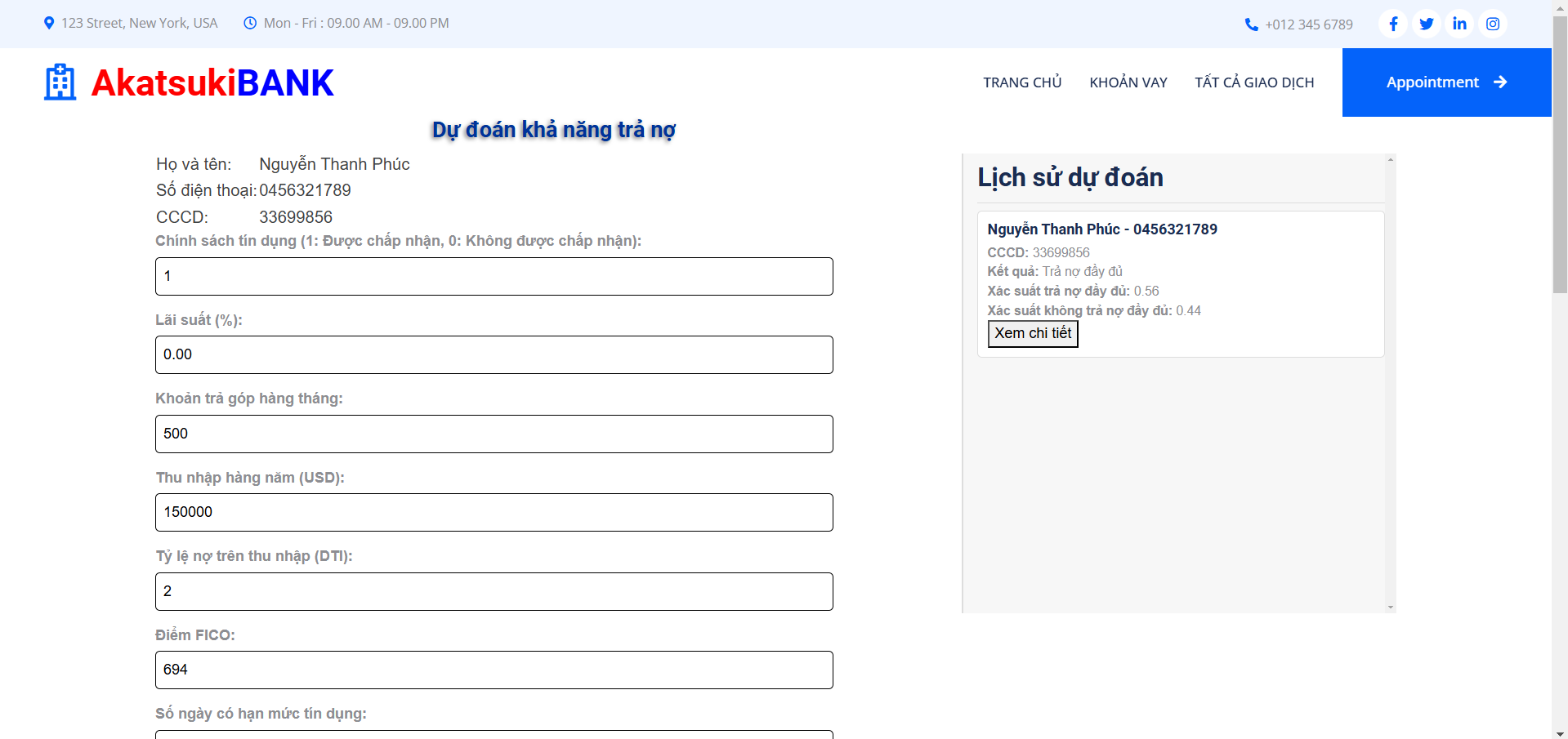
Hình 16: Các thông số cần nhập và thông tin đã nhập

Khi đã nhập xong, bấm kiểm tra sẽ hiện ra thông tin người đã nhập và những thông số cần thiết để dự đoán



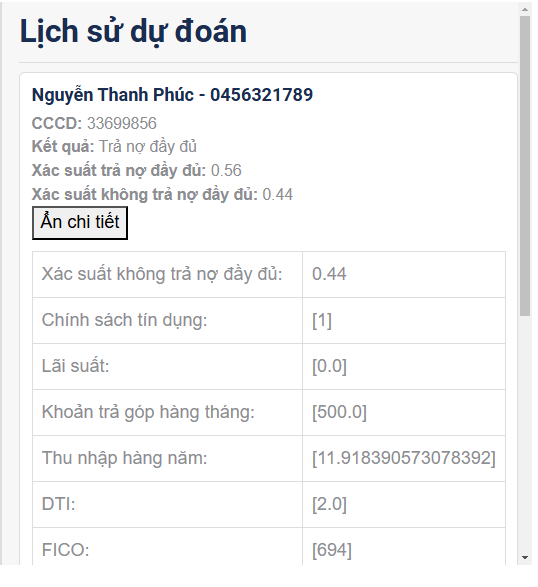
Hình 17: Kết quả dự đoán

Phần dự đoán sẽ được nằm bên dưới



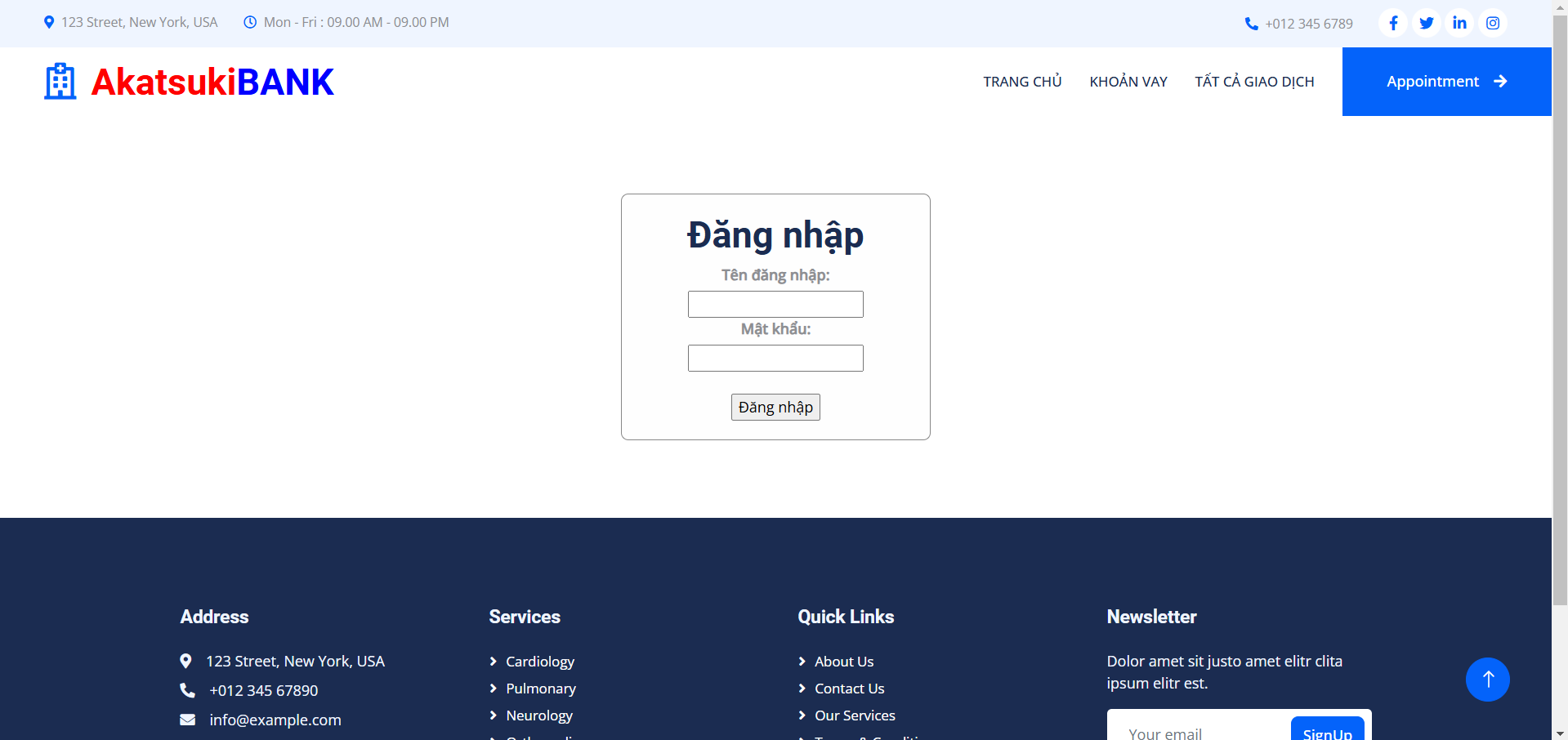
Hình 18: Hiển thị lên lịch sử

Phần lịch sử dự đoán của cá nhân người này cũng sẽ được hiển thị lên phần lịch sử dự đoán ở bên phải màn hình



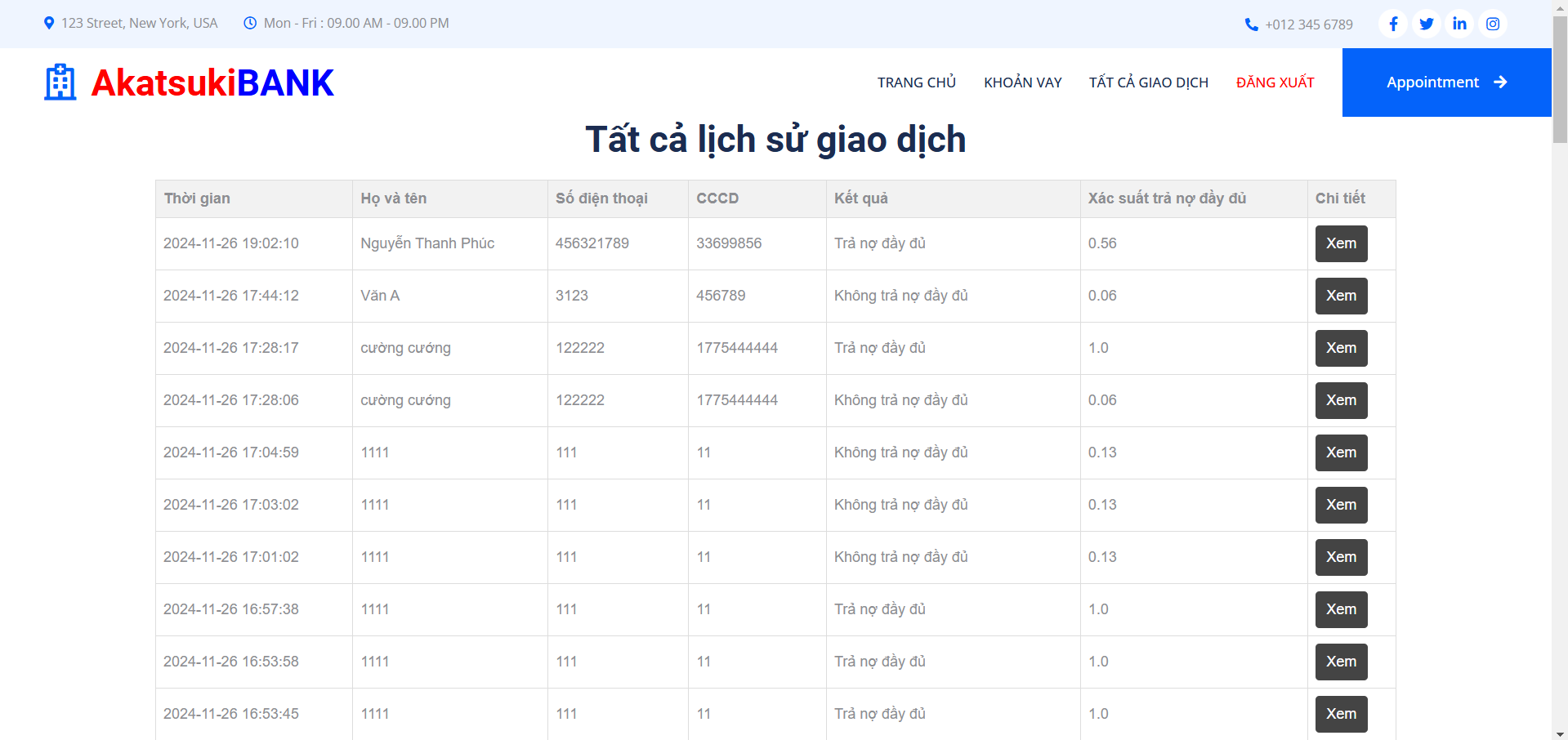
Hình 19: Xem lịch sử chi tiết

Bấm vào “ Xem chi tiết “ sẽ hiện ra chi tiết thông tin của người dùng kèm kết quả đã được dự đoán



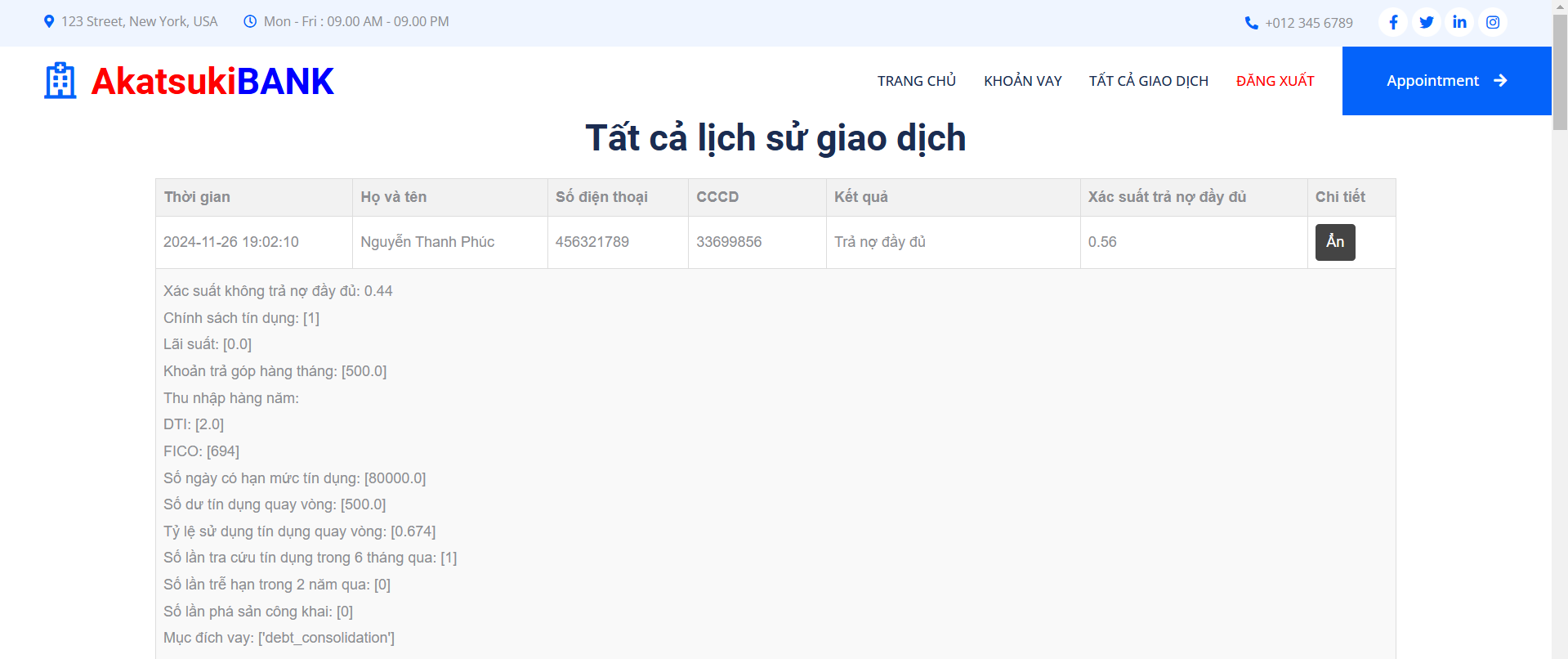
Hình 20: Giao diện đăng nhập admin

Muốn coi phần lịch sử tổng hợp thì bấm vào mục “ Tất cả giao dịch “ sẽ dẫn đến phần đăng nhập của admin



Hình 21: Tất cả lịch sử giao dịch

Tất cả lịch sử giao dịch sẽ được hiển thị ở đây



Hình 22: Xem chi tiết trong phần lịch sử giao dịch

Bấm xem chi tiết sẽ hiển thị tất cả thông tin người đã dự đoán

## 3. Kết quả thực nghiệm

Có thể thấy gần như kết quả dự đoán của mô hình Logistic Regression cho ra kết quả dự đoán rất tốt, tuy nhiên mức độ chính xác của mô hình vẫn chưa có thể đạt tới hoàn hảo.

Với mô hình Logistic Regression này thì có thể xác định được mức độ vay vốn và trả nợ của người dùng để có thể đưa ra kết quả có nên cho vay tiếp tục hay không với kết quả xấu hay kết quả tốt.

Mô hình Logistic này có độ chính xác thấp hơn không phải nhiều so với các mô hình khác nhưng nó đem lại sự hoàn chỉnh về dự đoán, tính ứng dụng rất cao trong các doanh nghiệp về vay vốn

# V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## 1. Kết luận

Dự án đã hoàn thành được mục tiêu đã được đề ra

- Mục tiêu: Dự đoán và đưa ra kết quả, xác suất trả nợ của khách hàng

Mô hình đạt mức độ chính xác cao, còn hạn chế về hiệu năng cũng như giao diện chưa được hoàn chỉnh 100% về ứng dụng.

Tuy nhiên mức độ dự đoán sai cũng có thể xay ra tùy thuộc vào thực tế có thay đổi những cục diện nào hay không

Tất cả các chỉ số tính toán độ chính xác còn lại cho thấy mô hình không overfit, underfit hoặc mang tính chất dự đoán cục bộ hay thể hiện sự mất cân bằng dữ liệu.

Tóm lại mô hình đã đạt được mục đích tính toán dự đoán khả năng trả nợ của khách hàng trong thị trường vay vốn của Việt Nam. Mô hình hoàn toàn có thể đưa vào thị trường thực tế vì cách sử dụng khá đơn giản và tính chính xác cao, đây cũng sẽ là một nền tảng vững chắc để tiếp tục phát triển thêm các mô hình dự đoán khác trong tương lai.

## 2. Hướng phát triển

- Về mô hình: Tối ưu hóa dữ liệu đầu vào, làm sạch dữ liệu, loại bỏ các giá trị ngoại lệ và xử lý dữ liệu bị thiếu. Sử dụng kỹ thuật chọn lọc đặc trưng (feature selection) để chọn ra những đặc trưng quan trọng nhất, loại bỏ những đặc trưng không liên quan hoặc gây nhiễu. Thử nghiệm với các giá trị khác nhau của tham số regularization (L1, L2) để giảm overfitting và cải thiện độ chính xác của mô hình. Sử dụng kỹ thuật Ensemble đểs kết hợp Logistic Regression với các mô hình khác, như Random Forest hoặc Gradient Boosting, nhằm cải thiện hiệu suất dự đoán. Sử dụng kỹ thuật bagging hoặc boosting để tăng cường khả năng dự đoán của mô hình.

- Về sản phẩm: Thiết kế tối ưu thêm giao diện sử dụng, thêm các tính năng khác như giao dịch cho vay vốn, đăng nhập khách hàng, thiết kế một ứng dụng có thể cho người sử dụng đăng nhập và thực hiện những giao dịch trực tuyến trên đó v.v…

# TÀI LIỆU THAM KHẢO

1. NV Quang năm 2022 ( <https://digital.lib.ueh.edu.vn/handle/UEH/67107> )
2. Tạp chí ngân hàng 2024 ( <https://digital.lib.ueh.edu.vn/handle/UEH/67107> )
3. Tạp chí công thương 2020 ( <https://tapchicongthuong.vn/du-doan-hanh-vi-tra-no-dung-han-cua-khach-hang-ca-nhan-tai-ngan-hang-thuong-mai-76060.htm> )
4. Y Liang ( <https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26644913.pdf>
5. E Oskarsson · 2021 ( <https://www.diva-portal.org/smash/get/diva2:1563211/FULLTEXT01.pdf> )
6. TN Dinh 2022 ( <https://ieeexplore.ieee.org/document/10068483> )