

ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN



BÁO CÁO THỰC HÀNH
MÔN: CƠ SỞ HẠ TẦNG CÔNG NGHỆ THÔNG TIN

ĐỀ TÀI

**MÔ HÌNH CHẤM ĐIỂM TÍN DỤNG ĐỂ TÌM KIẾM KHÁCH HÀNG
TIỀM NĂNG KẾT HỢP SỬ DỤNG MẠNG THẦN KINH NHÂN TẠO,
CÁC ĐƯỜNG HỒI QUY THÍCH ỨNG ĐA BIẾN VÀ PHÂN CỤM
K-MEANS**

GVHD: ThS. Nguyễn Thị Anh Thư

Nhóm sinh viên thực hiện: Nhóm 7

- | | |
|-------------------------|----------------|
| 1. Võ Nhất Phương | MSSV: 22521172 |
| 2. Lê Nguyễn Thùy Dương | MSSV: 22520298 |
| 3. Nguyễn Thế Tiến Đạt | MSSV: 22520225 |
| 4. Vũ Quang Huy | MSSV: 22520587 |
| 5. Đoàn Ngọc Thanh Sơn | MSSV: 21521385 |

TP. Hồ Chí Minh, tháng 4 năm 2024

MỤC LỤC

A. THÔNG TIN CHUNG	4
A1. Tên đề tài	4
A2. Thời gian thực hiện	4
A3. Tổng kinh phí	4
A4. Chủ nhiệm	4
A5. Thành viên đề tài	4
B. MÔ TẢ ĐỀ TÀI	5
B1. Giới thiệu về đề tài	5
B1.1 Tổng quan tình hình nghiên cứu thuộc lĩnh vực đề tài	5
B1.1.1 Ý tưởng đề tài	5
B1.1.2 Tính cấp thiết của đề tài	6
B1.1.3 Tính mới của đề tài	7
B1.2 Các dự án liên quan trong cùng lĩnh vực	8
B1.2.1 Dự án nghiên cứu về xác định khách hàng tiềm năng	8
B1.2.2 Dự án nghiên cứu về xây dựng mô hình chấm điểm tín dụng	9
B1.3 Mục tiêu thực hiện đề tài	10
B2. Mục tiêu, nội dung và phương pháp thực hiện	11
B3. Sản phẩm đề tài	30
B4. Kết quả dự kiến	30
B5. Đánh giá tính khả thi	30
C. KẾT LUẬN	31
D. TÓM TẮT NỘI DUNG	31
E. TÀI LIỆU THAM KHẢO	32

LỜI CẢM ƠN

Trong thời gian học tập và làm báo cáo thực hành đề tài “MÔ HÌNH CHẤM ĐIỂM TÍN DỤNG ĐỂ TÌM KIẾM KHÁCH HÀNG TIỀM NĂNG KẾT HỢP SỬ DỤNG MẠNG THẦN KINH NHÂN TẠO, CÁC ĐƯỜNG HỒI QUY THÍCH ỨNG ĐA BIẾN VÀ PHÂN CỤM K-MEANS” này, chúng em muốn gửi những lời cảm ơn và biết ơn chân thành nhất tới tất cả những người đã hỗ trợ, giúp đỡ nhóm về kiến thức và tinh thần trong quá trình thực hiện đồ án.

Trước hết, nhóm chúng em xin gửi lời cảm ơn tới cô Ths. Nguyễn Thị Anh Thư - Giảng viên Khoa Khoa học và Kỹ thuật Thông tin, Trường Đại học Công nghệ Thông tin ĐHQG HCM. Cô là người đã trực tiếp hướng dẫn, nhận xét, giúp đỡ chúng em trong suốt quá trình thực hiện đồ án.

Chúng em cũng xin gửi lời cảm ơn đến gia đình, bạn bè, người thân đã giúp đỡ động viên rất nhiều trong quá trình học tập và làm đồ án đúng thời gian quy định.

Do thời gian thực hiện có hạn, kiến thức còn nhiều hạn chế nên dự án thực hiện chắc chắn không tránh khỏi những thiếu sót nhất định. Chúng em rất mong nhận được ý kiến đóng góp của thầy cô giáo và các bạn để nhóm có thêm kinh nghiệm và tiếp tục hoàn thiện đồ án của mình.

Nhóm chúng em xin chân thành cảm ơn tất cả mọi người!

TP. Hồ Chí Minh, ngày 01 tháng 05 năm 2024

A. THÔNG TIN CHUNG

A1. Tên đề tài

- Tên tiếng Việt (IN HOA): MÔ HÌNH CHẤM ĐIỂM TÍN DỤNG ĐỂ TÌM KIẾM KHÁCH HÀNG TIỀM NĂNG KẾT HỢP SỬ DỤNG MẠNG THẦN KINH NHÂN TẠO, CÁC ĐƯỜNG HỒI QUY THÍCH ỨNG ĐA BIẾN VÀ PHÂN CỤM K-MEANS
- Tên tiếng Anh (IN HOA): CREDIT SCORING MODEL FOR FINDING POTENTIAL CUSTOMERS THAT COMBINES THE USE OF ARTIFICIAL NEURAL NETWORKS, MULTIVARIATE ADAPTIVE REGRESSION SPLINES AND K-MEANS CLUSTERING

A2. Thời gian thực hiện

12 tháng (kể từ khi được duyệt)

A3. Tổng kinh phí

Ước tính: 200 triệu đồng

A4. Chủ nhiệm

Họ và tên: Võ Nhất Phương

Ngày, tháng năm sinh: 13/03/2004

Giới tính (Nam/Nữ): Nữ

MSSV: 22521172

Số điện thoại liên lạc: 0365486141

Đơn vị (Khoa): Khoa học và Kỹ thuật Thông tin

A5. Thành viên đề tài

TT	Họ tên	MSSV	Khoa
1	Lê Nguyễn Thùy Dương	22520298	Khoa học và Kỹ thuật Thông tin

2	Nguyễn Thế Tiến Đạt	22520225	Khoa học và Kỹ thuật Thông tin
3	Vũ Quang Huy	22520587	Khoa học và Kỹ thuật Thông tin
4	Đoàn Ngọc Thanh Sơn	21521385	Khoa học và Kỹ thuật Thông tin

B. MÔ TẢ ĐỀ TÀI

B1. Giới thiệu về đề tài

B1.1 Tổng quan tình hình nghiên cứu thuộc lĩnh vực đề tài

B1.1.1 Ý tưởng đề tài

Trong bối cảnh phát triển của nền kinh tế số, việc áp dụng công nghệ thông tin vào quản lý tín dụng là một xu hướng không thể phủ nhận, đặc biệt là ở Việt Nam. Hiện nay, việc chấm điểm tín dụng cá nhân đã trở thành một phần quan trọng trong quy trình xác định khả năng thanh toán của khách hàng, không chỉ giúp các tổ chức tín dụng đánh giá rủi ro mà còn tạo điều kiện thuận lợi cho người dân tiếp cận các dịch vụ tài chính.

Trong thực tế, quá trình chấm điểm tín dụng cá nhân ở Việt Nam vẫn đang trong giai đoạn phát triển với sự tham gia tích cực của các tổ chức tín dụng, cũng như các doanh nghiệp công nghệ. Các thuật toán và mô hình chấm điểm tín dụng được ngày càng phát triển và tinh chỉnh để phản ánh chính xác hơn khả năng thanh toán của cá nhân. Tuy nhiên, còn nhiều thách thức cần vượt qua như việc tích hợp dữ liệu từ nhiều nguồn khác nhau, đảm bảo tính minh bạch và bảo mật thông tin cá nhân.

Với tình hình thị trường đang dần chuyển biến từ mô hình truyền thống sang mô hình số, việc áp dụng giải pháp chấm điểm tín dụng cá nhân sử dụng các thuật toán dữ liệu lớn đang là một xu thế tất yếu. Ứng dụng Big data không chỉ giúp tối ưu hóa quy trình đánh giá tín nhiệm mà còn mở ra cơ hội tiềm năng trong việc tăng cường tiếp cận, phục vụ cho đa dạng nhu cầu tài chính của người dân và doanh nghiệp.

Vì thế sau khi khảo sát Cơ sở hạ tầng Công nghệ thông tin của Công ty TNHH MTV Viễn thông Quốc tế FPT Hồ Chí Minh, nhóm chúng em nhận thấy rằng:

- Công ty TNHH MTV Viễn thông Quốc tế FPT Hồ Chí Minh đã có nền tảng và kỹ thuật hiện đại trong lĩnh vực Công nghệ thông tin, giúp dễ dàng tích hợp và triển khai các giải pháp chấm điểm tín dụng mới.
- Công ty TNHH MTV Viễn thông Quốc tế FPT Hồ Chí Minh cung cấp một loạt dịch vụ với mức giá đa dạng, từ các dịch vụ phổ thông đến các gói dịch vụ cao cấp. Với những dịch vụ có giá thành cao, có một số khách hàng có nhu cầu sử dụng nhưng không có khả năng thanh toán ngay lập tức toàn bộ số tiền mua hàng. Do đó, việc áp dụng giải pháp chấm điểm tín dụng cá nhân giúp Công ty xác định và đánh giá khả năng thanh toán của từng khách hàng để tìm ra khách hàng tiềm năng, từ đó quyết định việc cho phép trả góp cho đối tượng khách hàng nào sẽ là hợp lý và an toàn cho Công ty.

Vì thế nhóm chúng em đề xuất mô hình chấm điểm tín dụng sử dụng các thuật toán dữ liệu lớn.

B1.1.2 Tính cấp thiết của đề tài

Trong thời đại hiện nay, vấn đề về tìm kiếm và thu hút khách hàng tiềm năng đối với các doanh nghiệp, đặc biệt là các công ty về công nghệ trở nên cực kỳ cấp thiết. Trong bối cảnh cạnh tranh ngày càng gay gắt và sự thay đổi nhanh chóng của thị trường, việc hiểu rõ và đáp ứng đúng đắn nhu cầu của khách hàng là yếu tố quyết định sự thành công của một doanh nghiệp.

Việc áp dụng giải pháp chấm điểm tín dụng trong tìm kiếm và xác định khách hàng tiềm năng không chỉ giúp doanh nghiệp hiểu rõ hơn về khách hàng mục tiêu mà còn tăng cơ hội thành công trong việc thu hút và giữ chân khách hàng. Bên cạnh đó còn giúp doanh nghiệp xác định và đánh giá rõ ràng khả năng thanh toán của từng khách hàng, giúp quản lý rủi ro tốt hơn, giảm thiểu rủi ro nợ xấu và tăng cường tính bền vững cho doanh nghiệp.

Quá trình nghiên cứu về cơ sở hạ tầng công nghệ thông tin tại Công ty TNHH MTV Viễn thông Quốc tế FPT cho thấy rằng công ty vẫn chưa áp dụng các công nghệ dữ liệu lớn

tiên tiến trong việc chấm điểm tín dụng để tìm kiếm khách hàng tiềm năng. Khi không có hệ thống chấm điểm tín dụng hiệu quả, việc xác định và đánh giá khách hàng tiềm năng trở nên khó khăn hơn, gây ra những thách thức trong quản lý và kiểm soát, đặc biệt là khi xảy ra các vấn đề liên quan đến thông tin khách hàng và quản lý tín dụng. Ngoài ra, thiếu minh bạch và chính xác trong thông tin tín dụng có thể dẫn đến sự mất niềm tin của khách hàng và ảnh hưởng đến hình ảnh của công ty.

Để giải quyết những thách thức này, đề tài đề xuất sử dụng kết hợp giữa mạng thần kinh nhân tạo, đường hồi quy thích ứng đa biến và phương pháp K-means trong việc chấm điểm tín dụng nhằm xác định khách hàng tiềm năng. Giải pháp giúp các bên liên quan trong quá trình quản lý thông tin tín dụng có thể đưa ra quyết định chính xác, nhanh chóng về việc cung cấp dịch vụ tín dụng hay không, từ đó tăng cơ hội thành công và giảm thiểu rủi ro cho công ty.

B1.1.3 Tính mới của đề tài

- Tạo ra một mô hình chấm điểm tín dụng mạnh mẽ và linh hoạt: Sử dụng sức mạnh của mạng nơ-ron nhân tạo (ANNs) và đường hồi quy thích ứng đa biến (MARS) để xây dựng mô hình chấm điểm tín dụng. Các mô hình này sẽ phân tích một loạt các biến như lịch sử tín dụng, thu nhập và dư nợ để dự đoán khả năng trả nợ của mỗi cá nhân hoặc tổ chức. Kết quả từ hai phương pháp này sẽ cung cấp điểm số tín dụng chính xác và đáng tin cậy, giúp cải thiện quá trình ra quyết định về tín dụng
- Phân loại khách hàng theo tiềm năng: Sử dụng thuật toán K-means để phân cụm khách hàng thành các nhóm dựa trên điểm số tín dụng. Các nhóm này sẽ phản ánh mức độ rủi ro và tiềm năng trả góp của từng khách hàng.
- Nâng cao quản lý rủi ro và hiệu suất kinh doanh: Bằng cách kết hợp các mô hình chấm điểm tín dụng với phân loại khách hàng, công ty TNHH MTV Viễn thông Quốc tế FPT có thể đánh giá và quản lý rủi ro một cách hiệu quả hơn. Đồng thời

tối ưu hóa chiến lược kinh doanh bằng cách tập trung vào nhóm khách hàng có tiềm năng cao nhất, từ đó tăng cường hiệu suất kinh doanh và lợi nhuận.

B1.2 Các dự án liên quan trong cùng lĩnh vực

Chấm điểm tín dụng nhằm đánh giá rủi ro tín dụng đóng vai trò quan trọng trong việc hỗ trợ chính xác các tổ chức tài chính hay các bộ phận tài chính trong công ty xác định chính sách ngân hàng và chiến lược thương mại của mình. Bài toán chấm điểm tín dụng thường được mô hình hóa dưới dạng bài toán nhị phân trên cơ sở trả nợ và do đó có thể khai thác các kỹ thuật machine learning phù hợp [1].

Đã có nhiều dự án khác nhau ứng dụng thuật toán dữ liệu lớn được tiến hành trong lĩnh vực chấm điểm tín dụng tại các công ty. Do đó việc tìm hiểu các công trình nghiên cứu liên quan đến đề tài sẽ cho người đọc cái nhìn tổng quan cũng như hiểu biết sâu sắc về đề tài. Từ đó có thể lựa chọn các giải pháp phù hợp, đồng thời đề xuất phương pháp cải tiến có thể mang lại hiệu quả tốt cho kết quả tìm ra khách hàng tiềm năng dựa trên mô hình chấm điểm tín dụng.

B1.2.1 Dự án nghiên cứu về xác định khách hàng tiềm năng

K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data (Phương pháp phân cụm K-Means để phân khúc khách hàng thông minh bằng cách sử dụng dữ liệu hành vi mua hàng của khách hàng) [2]

- Mục đích của việc phân khúc khách hàng là xác định cách thức ứng xử với khách hàng trong từng hạng mục nhằm tăng lợi nhuận của từng khách hàng cho doanh nghiệp. Phân khúc khách hàng hỗ trợ doanh nghiệp xác định khách hàng có lợi nhuận để đáp ứng nhu cầu của họ bằng cách tối ưu hóa các dịch vụ và sản phẩm.
- Có một số loại yếu tố phân khúc khách hàng là tâm lý nhân khẩu học, hành vi và địa lý. Trong nghiên cứu này, yếu tố hành vi của khách hàng được chú trọng. Do đó, người dùng sẽ được phân tích bằng thuật toán phân cụm để xác định hành vi mua hàng của hệ thống Thương mại điện tử.

- Để xử lý dữ liệu được thu thập và phân khúc khách hàng, một thuật toán học được gọi là phân cụm K-Means được sử dụng. Phân cụm K-Means được triển khai để giải quyết các vấn đề về phân cụm.
- Ưu điểm:
 - + Hiệu suất thuật toán đã chọn được hiển thị trên dữ liệu lịch sử mua hàng của khách hàng để phân khúc khách hàng cao.
 - + Mức độ xác thực được cải thiện đáng kể, chẳng hạn như các giá trị danh nghĩa được làm mịn bằng điểm z.
 - + Giúp các nhà cung cấp xác định và tập trung vào phân khúc có lợi nhuận cao đến phân khúc có lợi nhuận thấp nhất có thể đóng vai trò quan trọng trong việc cải thiện hoạt động kinh doanh.
 - + Có thể được sử dụng trên bất kỳ bộ dữ liệu nào liên quan đến kinh doanh.

B1.2.2 Dự án nghiên cứu về xây dựng mô hình chấm điểm tín dụng

A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines (Mô hình chấm điểm tín dụng kết hợp hai giai đoạn sử dụng mạng thần kinh nhân tạo và các đường hồi quy thích ứng đa biến) [3]

- Cơ sở lý luận của các phân tích trước hết là sử dụng MARS trong việc xây dựng mô hình chấm điểm tín dụng, các biến quan trọng thu được sau đó được dùng làm nút đầu vào của mô hình mạng nơ ron.
- Để chứng minh tính hiệu quả và tính khả thi của quy trình lập mô hình được đề xuất, các nhiệm vụ chấm điểm tín dụng được thực hiện trên một tập dữ liệu cho vay mua nhà của ngân hàng bằng cách sử dụng phương pháp xác thực chéo.
- Ưu điểm: Theo kết quả cho thấy, phương pháp kết hợp được đề xuất vượt trội hơn so với kết quả sử dụng phân tích phân biệt tuyến tính (LDA), hồi quy logistic, mạng lưới nơ ron nhân tạo và MARS.
 - + Mạng nơ ron nhân tạo cung cấp một giải pháp thay thế mới cho LDA và hồi quy logistic trong việc xử lý các nhiệm vụ chấm điểm tín dụng, đặc biệt trong các tình huống mà các biến phụ thuộc và biến độc lập thể hiện mối

quan hệ phi tuyến tính phức tạp. Nhưng quá trình đào tạo kéo dài để có được cấu trúc liên kết mạng tối ưu, không dễ xác định tầm quan trọng tương đối của các biến đầu vào tiềm năng và những khó khăn diễn giải nhất định. Do đó đã hạn chế khả năng ứng dụng của nó trong việc xử lý các vấn đề phân loại chung và chấm điểm tín dụng.

- + Spline hồi quy thích ứng đa biến (MARS) không có nhược điểm của LDA và hồi quy logistic, MARS có khả năng mô hình hóa mối quan hệ phi tuyến tính phức tạp giữa các biến mà không cần các giả định mô hình mạnh mẽ. Mặt khác, không giống như mạng nơ-ron, MARS có thể nắm bắt được tầm quan trọng tương đối của các biến độc lập đối với biến phụ thuộc khi xem xét nhiều biến độc lập tiềm năng. Thứ ba, MARS không cần quá trình đào tạo dài và do đó có thể tiết kiệm rất nhiều thời gian xây dựng mô hình, đặc biệt khi tập dữ liệu rất lớn. Cuối cùng, một ưu điểm mạnh mẽ của MARS so với các kỹ thuật phân loại khác là mô hình kết quả có thể được diễn giải dễ dàng.

=> MARS sẽ là một công cụ hỗ trợ tốt cho mạng nơ-ron nhân tạo

B1.3 Mục tiêu thực hiện đề tài

- Mục tiêu thực hiện đề tài nhằm tạo ra một ứng dụng BA Chấm điểm tín dụng an toàn và minh bạch. Ứng dụng dành cho nhân viên Bộ phận Marketing và bán hàng của Công ty TNHH MTV Viễn thông Quốc tế FPT sử dụng với mục đích xác định khách hàng tiềm năng trên mô hình chấm điểm tín dụng. Từ đó, Bộ phận Marketing và bán hàng sẽ đưa ra quyết định cho phép khách hàng trả góp hay là không.
- Ứng dụng Chấm điểm tín dụng có các tính năng cụ thể như sau:
 - + Thu thập dữ liệu: Thu thập dữ liệu từ nhiều nguồn khác nhau bao gồm lịch sử tín dụng của khách hàng, thông tin tài chính và dữ liệu cá nhân khác.
 - + Phân tích dữ liệu: Sử dụng các thuật toán phân tích để đánh giá rủi ro tín dụng và khả năng trả góp của khách hàng dựa trên dữ liệu được thu thập.

- + Chấm điểm tín dụng: Tổng hợp thông tin từ quá trình phân tích dữ liệu để tạo ra một điểm số tín dụng của khách hàng. Điểm số này được sử dụng để quyết định liệu khách hàng có đủ điều kiện để trả góp hay không.
- + Giao diện người dùng: Phát triển một giao diện người dùng thân thiện để các nhân viên trong Bộ phận Marketing và bán hàng có thể sử dụng. Giao diện này cần cung cấp các tính năng như nhập thông tin khách hàng, xem kết quả chấm điểm tín dụng và đưa ra quyết định về việc cho phép trả góp hay không.
- + Bảo mật và tuân thủ quy định: Ứng dụng tuân thủ các quy định về bảo mật dữ liệu và quyền riêng tư, đặc biệt là khi làm việc với thông tin cá nhân của khách hàng.

B2. Mục tiêu, nội dung và phương pháp thực hiện

Nội dung 1: Cơ sở lý thuyết của việc ứng dụng thuật toán Big Data Mạng nơ ron nhân tạo (ANN), thuật toán Splines hồi quy thích ứng đa biến (MARS) và thuật toán Phân cụm K-means

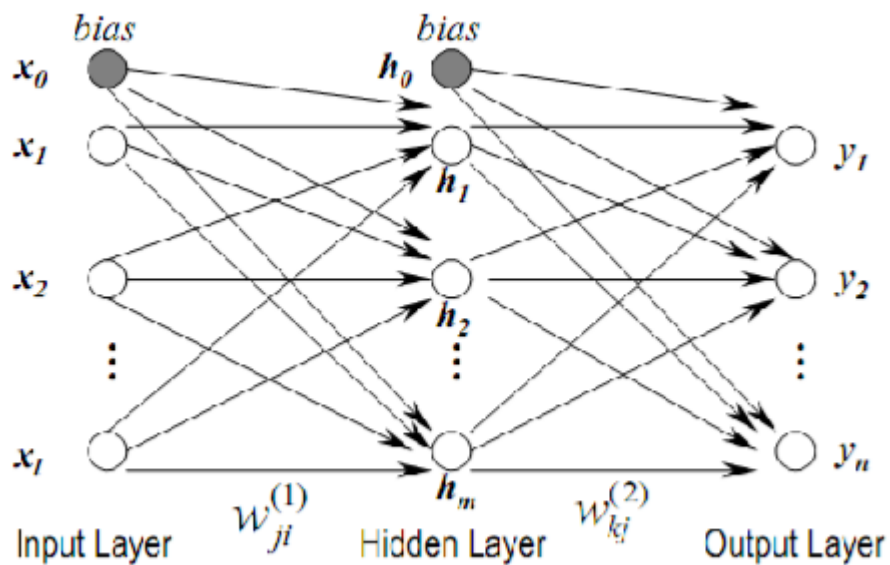
Mục tiêu 1: Nắm vững cơ sở lý thuyết của các thuật toán Big Data như Mạng nơ ron nhân tạo (ANN), Splines hồi quy thích ứng đa biến (MARS) và Phân cụm K-means để áp dụng chúng trong việc xây dựng ứng dụng Chấm điểm tín dụng.

Phương pháp 1:

Thuật toán Mạng nơ ron nhân tạo (ANNs): Mạng nơ ron truyền thẳng (MLP)

Đây là một chuỗi những thuật toán được đưa ra để tìm kiếm các mối quan hệ cơ bản trong tập hợp các dữ liệu. Thông qua việc bắt chước cách thức hoạt động từ não bộ con người. Nói cách khác, mạng nơ ron nhân tạo được xem là hệ thống của các tế bào thần kinh nhân tạo. Trong lĩnh vực tài chính, mạng nơ ron nhân tạo hỗ trợ cho quá trình phát triển các quy trình như: giao dịch thuật toán, dự báo chuỗi thời gian, phân loại chứng khoán, mô hình rủi ro tín dụng và xây dựng chỉ báo độc quyền và công cụ phát sinh giá.

Mạng nơ ron truyền thẳng (Feedforward), hay còn gọi là mạng perceptron nhiều lớp (MLP - Multi-layer perceptrons) là mạng nơ ron phổ biến nhất. Chúng bao gồm một lớp đầu vào, một hoặc nhiều lớp ẩn và một lớp đầu ra. Mạng nơ ron truyền thẳng xử lý dữ liệu một chiều, từ lớp đầu vào đến lớp đầu ra. Mỗi nút trong cùng một lớp được kết nối với tất cả các nút trong lớp tiếp theo. Mạng truyền thẳng sử dụng một quy trình phản hồi để cải thiện dự đoán theo thời gian [4].



Thuật toán Splines hồi quy thích ứng đa biến (MARS):

Multivariate Adaptive Regression Splines (MARS) là một phương pháp trong học máy và thống kê được sử dụng cho việc dự đoán và mô hình hóa mối quan hệ giữa biến phụ thuộc và biến độc lập trong các tập dữ liệu có tính phi tuyến tính và đa chiều. MARS kết hợp sự linh hoạt của một số đặc trưng cơ bản của hồi quy tuyến tính với khả năng tạo ra các đường cong phi tuyến tính thông qua việc sử dụng các spline functions.

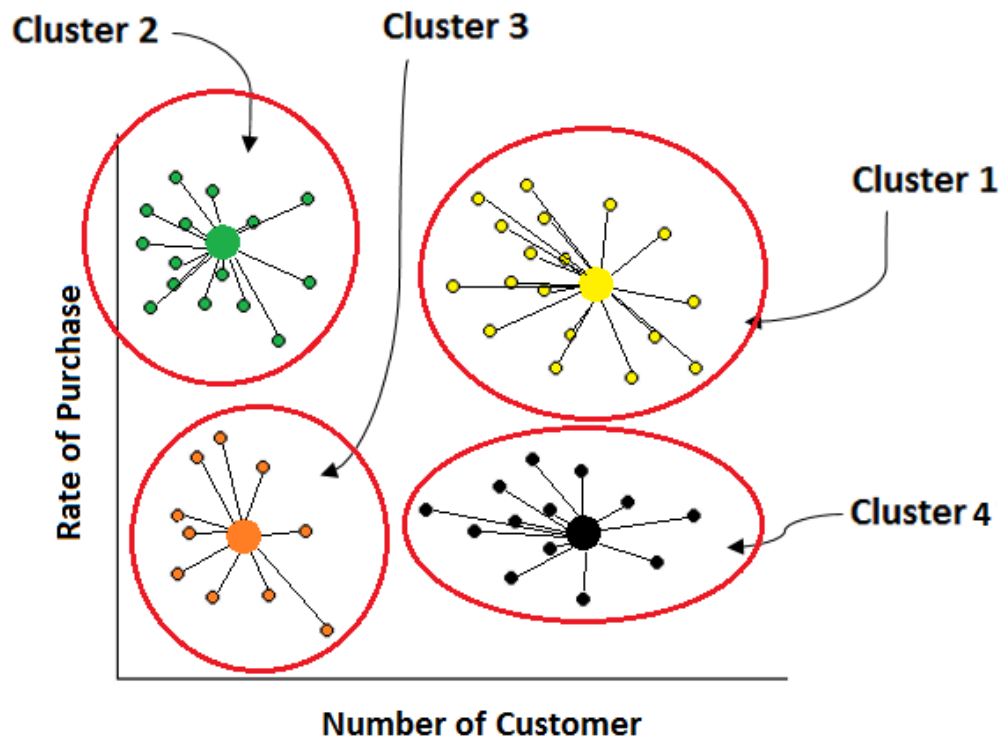
Một trong những ưu điểm của MARS là khả năng xử lý dữ liệu có nhiều biến và tương tác phức tạp một cách hiệu quả. Nó cũng có khả năng tự động loại bỏ các biến không quan trọng và xác định các điểm cắt quan trọng trong dữ liệu. Điều này làm cho MARS

trở thành một công cụ hữu ích cho việc khám phá và mô hình hóa dữ liệu phi tuyến tính trong nhiều ứng dụng thực tế.

Thuật toán Phân cụm K-means:

K-mean clustering là một phương pháp để tìm các cụm và hạt nhân - trung tâm của cụm trong một tập hợp dữ liệu không được gắn nhãn. Người ta chọn số lượng hạt nhân cụm mong muốn phân chẳng hạn như k cụm. Thuật toán K-mean di chuyển lặp đi lặp lại các hạt nhân để giảm thiểu tổng số trong phương sai cụm. Với một tập hợp các hạt nhân ban đầu, thuật toán Kmeans lặp lại hai bước:

- Đối với mỗi hạt nhân, tính toán khoảng cách giữa các training point với nó và nếu gần nó hơn -> sẽ gán là cụm của hạt nhân đấy.
- Sau khi phân được cụm như ở bước trên, thì tiếp theo các training point của các cụm tính toán vector trung bình để được vị trí của hạt nhân mới và lặp lại bước trên đến khi không thể thay đổi được vị trí hạt nhân nữa [5].



Kết quả 1:

Sẵn sàng để áp dụng các thuật toán Big Data kể trên vào việc xây dựng ứng dụng Chấm điểm tín dụng.

Nội dung 2: Sử dụng thuật toán Mạng nơ ron nhân tạo (ANN), Splines hồi quy thích ứng đa biến (MARS) để xây dựng mô hình chấm điểm tín dụng

Mục tiêu 2: Xây dựng một mô hình chấm điểm tín dụng có độ chính xác cao và khả năng dự đoán tốt trên các dữ liệu thông tin về khách hàng.

Phương pháp 2:

B1: Thu thập dữ liệu:

Để kiểm chứng mô hình đề xuất, nghiên cứu sử dụng dữ liệu thu thập từ nhiều nguồn khác nhau:

- Thông tin khách hàng: Họ tên, tuổi, giới tính, trình độ học vấn và tình trạng hôn nhân của khách hàng
- Tình trạng nghề nghiệp: Khách hàng có phải là người sống phụ thuộc, thất nghiệp, làm việc bán thời gian hay có việc làm ổn định
- Thu nhập hàng tháng của khách hàng
- Khách hàng có các tài khoản thanh toán hay tài khoản tiết kiệm hay không
- Sở hữu nhà: Yếu tố này có thể ảnh hưởng đến tính ổn định của khách hàng và khả năng thanh toán của khách hàng
- Lịch sử tín dụng: Thu thập thông tin thanh toán và lịch sử giao dịch từ các đơn đặt hàng được đặt tại công ty FPT. Thông tin này bao gồm các giao dịch thanh toán bằng thẻ tín dụng/ghi nợ hoặc các phương thức thanh toán trực tuyến khác như ví điện tử. Ngoài ra còn có thể yêu cầu khách hàng cung cấp thông tin thêm về các khoản nợ hiện tại, nợ xấu (nếu có).
- Các thông tin khác có thể có như: số lượng người trong gia đình, chi phí sinh hoạt,....

B2: Tiền xử lý dữ liệu

Các bước thực hiện:

- Tiêu chuẩn hóa dữ liệu: Đối với các thuộc tính số học như thu nhập hàng tháng có thể cần tiêu chuẩn hóa dữ liệu để đảm bảo rằng các biến có phạm vi giá trị tương tự giúp mô hình học được hiệu quả hơn.
- Xử lý dữ liệu thiếu: Kiểm tra và xử lý dữ liệu thiếu nếu có. Có thể sử dụng các phương pháp như điền giá trị trung bình, median hoặc sử dụng mô hình dự đoán để dự đoán các giá trị bị thiếu.

- Biến đổi dữ liệu: Có thể áp dụng các biến đổi dữ liệu như chuyển đổi logarit, căn bậc hai hoặc các biến đổi khác để làm cho phân phối của dữ liệu gần gũi với phân phối chuẩn giúp mô hình hoạt động tốt hơn.
- Loại bỏ nhiễu: Kiểm tra và loại bỏ các giá trị ngoại lệ hoặc nhiễu trong dữ liệu. Các giá trị ngoại lệ có thể ảnh hưởng đến khả năng dự đoán của mô hình.
- Phân tích đặc điểm của dữ liệu: Sử dụng các phương pháp như phân tích tương quan, phân phối và biểu đồ để hiểu rõ hơn về tính phân bố của các thuộc tính và mối quan hệ giữa chúng.
- Tách tập dữ liệu: Tách dữ liệu thành tập huấn luyện và tập kiểm tra để đánh giá hiệu suất của mô hình trên dữ liệu mới.
- Chuẩn bị dữ liệu cho mô hình: Đảm bảo rằng dữ liệu đã được chuẩn bị một cách thích hợp để sử dụng trong mô hình. Cần chuyển đổi các biến phân loại thành dạng số học nếu cần thiết và chuẩn bị các biến đầu vào, biến mục tiêu cho mô hình.

B3: Lựa chọn thuật toán phân loại:

Thuật toán Splines hồi quy thích ứng đa biến (MARS): MARS giúp xác định các biến quan trọng trong mô hình bằng cách xem xét tầm quan trọng tương đối của mỗi biến đối với biến mục tiêu. Sau đó MARS tự động loại bỏ các biến không quan trọng, chọn ra các biến quan trọng làm nút đầu vào cho mô hình Mạng nơ ron nhân tạo.

Thuật toán Mạng nơ ron nhân tạo (ANNs): Mạng nơ ron truyền thẳng

- Mạng nơ ron nhân tạo truyền thẳng (MLP) bao gồm ít nhất ba tầng:
 - + Tầng đầu vào (input layer): Tầng này nhận các giá trị đầu vào từ dữ liệu và chuyển tiếp chúng đến các tầng tiếp theo. Các nút trong tầng này không phải là nơ-ron, mà chỉ là đại diện cho dữ liệu.
 - + Tầng ẩn (hidden layer): Tầng này bao gồm các nơ-ron nhân tạo, được kết nối với các tầng trước và sau. Các nơ-ron trong tầng này thực hiện các phép

tính và học hỏi từ dữ liệu. Một mạng nơ-ron có thể có nhiều tầng ẩn, tùy thuộc vào độ phức tạp của bài toán.

- + Tầng đầu ra (output layer): Tầng này nhận các giá trị từ tầng ẩn cuối cùng và cho ra kết quả đầu ra của mạng. Các nơ-ron trong tầng này thường có chức năng kích hoạt khác với các tầng ẩn, để phù hợp với loại bài toán cần giải quyết [6].
- Nghiên cứu sử dụng mô hình ANNs gồm 9 neural đầu vào, 1 neural phát tín hiệu đầu ra.

Ký hiệu và tên biến		Cách đo lường
X1	Họ tên	
X2	Tuổi	Nhận các giá trị: (1) Từ 18-25 tuổi ; (2) Từ 26-40 tuổi; (3) Từ 41-60 tuổi ; (4) Trên 60 tuổi
X3	Giới tính	Nhận các giá trị: (1) Nữ ; (2) Nam ; (3) Khác
X4	Trình độ học vấn	Nhận các giá trị: (1) Dưới trung học; (2) Trung học; (3) Cao đẳng/ Đại học; (4) Sau đại học
X5	Tình trạng hôn nhân	Nhận các giá trị: (1) Đã kết hôn ; (2) Độc thân;
X6	Tình trạng nghề nghiệp	Nhận các giá trị: (1) Thất nghiệp ; (2) Có công việc ổn định; (3) Làm việc bán thời gian

X7	Mức thu nhập ròng hàng tháng	Nhận các giá trị (1) Dưới 3 triệu đồng/tháng (2) Từ 3-8 triệu đồng/tháng (3) > 8-15 triệu đồng/tháng (4) > 15 triệu đồng/tháng
X8	Tình trạng sở hữu nhà	Nhận các giá trị (1) Khác; (2) Ở chung với gia đình; (3) Thuê nhà ; (4) Chủ sở hữu
X9	Thông tin thanh toán, giao dịch	Nhận các giá trị: Tổng số tiền thanh toán, Thời gian thanh toán, Phương thức thanh toán
Y	Số điểm tín dụng	

B4: Huấn luyện thuật toán phân loại:

Dữ liệu sử dụng trong mô hình được chia thành ba phần:

- Dữ liệu dùng để huấn luyện
- Dữ liệu dùng để kiểm tra tính khái quát của mô hình
- Dữ liệu dùng để kiểm định mô hình ngoài mẫu

Mô hình tổng thể được thể hiện như sau:

$$Y = X_1W_1 + b_1 + X_2W_2 + b_2 + X_3W_3 + b_3 + X_4W_4 + b_4 + X_5W_5 + b_5 + X_6W_6 + b_6 + X_7W_7 + b_7 + X_8W_8 + b_8 + X_9W_9 + b_9$$

Trong đó: W_i là trọng số liên kết do mô hình tự khởi tạo; b_i là độ lệch do chương trình tự điều chỉnh cho phù hợp mô hình.

Quá trình huấn luyện:

Dưới đây là quá trình phát triển chi tiết từng bước trong việc huấn luyện mô hình chấm điểm tin dụng sử dụng thuật toán MLP:

- Khởi tạo trọng số và thiên vị:
 - + Trọng số (weights) của mạng nơ-ron cần được khởi tạo ban đầu. Các trọng số này được chọn ngẫu nhiên từ một phân phối chuẩn hoặc một phân phối đồng đều trong một khoảng nhất định.
 - + Thiên vị (biases) cũng cần được khởi tạo. Như trọng số, các thiên vị có thể được chọn ngẫu nhiên từ một phân phối chuẩn hoặc một phân phối đồng đều.
- Truyền dữ liệu:
 - + Dữ liệu đầu vào: Thông tin 9 neurals đầu vào được cung cấp cho mạng nơ-ron.
 - + Trong quá trình lan truyền tiến (forward propagation), dữ liệu được truyền qua các lớp của mạng nơ-ron từ lớp đầu vào đến lớp đầu ra.
 - + Trong mỗi nơ-ron, giá trị đầu ra được tính toán bằng cách áp dụng hàm kích hoạt Sigmoid cho tổng trọng số của đầu vào và thiên vị.
- Tính toán lỗi: Sau khi dữ liệu đã được truyền qua mạng nơ-ron và đầu ra được tính toán, lỗi cần được tính toán để đo lường sự khác biệt giữa đầu ra dự đoán và đầu ra thực tế. Hàm mất mát (loss function) thường được sử dụng để tính toán lỗi. Ví dụ: Mean Squared Error (MSE), Cross-Entropy Loss.
- Cập nhật trọng số:
 - + Sau khi lỗi đã được tính toán, backpropagation được sử dụng để tính toán đạo hàm của hàm mất mát đối với các trọng số của mạng nơ-ron.
 - + Gradient Descent được áp dụng để điều chỉnh các trọng số theo hướng và mức độ được xác định bởi đạo hàm.
 - + Một tỷ lệ học (learning rate) cũng cần được xác định để điều chỉnh mức độ cập nhật trọng số.

- Lặp lại: Quá trình trên được lặp lại nhiều lần từ bước 2,3,4 thông qua tập dữ liệu huấn luyện (training dataset). Mỗi lần lặp lại một lần gọi là một epoch. Số lượng epoch có thể được xác định trước hoặc được dừng lại khi độ chính xác của mô hình trên tập dữ liệu kiểm tra không còn tăng nữa. Quá trình này được lặp lại cho đến khi mô hình đạt được hiệu suất mong muốn trên tập dữ liệu huấn luyện hoặc cho đến khi kết thúc số lượng epoch được chỉ định.

Kết quả 2:

- Kết quả phân loại tổng thể cho thấy rõ rằng phương pháp kết hợp cả thuật toán Mạng nơ ron nhân tạo và Splines hồi quy thích ứng đa biến đem lại kết quả dự báo tốt
- Output của mô hình chấm điểm tín dụng sẽ là số điểm tín dụng của khách hàng tại Công ty TNHH MTV Viễn thông quốc tế FPT

Phần mềm sử dụng:

- Ngôn ngữ lập trình: Python
- Môi trường phát triển: Jupyter Notebook
- Các gói sử dụng:
 - + pandas: pandas được sử dụng cho việc xử lý và làm sạch dữ liệu.
 - + numpy: numpy được sử dụng cho các phép toán toán học và manipulation trên mảng.
 - + seaborn và matplotlib.pyplot: seaborn và matplotlib.pyplot được sử dụng để trực quan hóa dữ liệu.
 - + scikit plot: scikit plot được sử dụng để trực quan hóa các chỉ số đánh giá của mô hình học máy.
 - + sklearn: sklearn (Scikit-learn) là một thư viện mạnh mẽ cho học máy trong Python. Nó cung cấp một loạt các thuật toán học máy và các phương pháp đánh giá mô hình.

- + shap: shap được sử dụng để trực quan hóa sự quan trọng của các đặc trưng trong mô hình.

Nội dung 3: Sử dụng điểm tín dụng khách hàng từ mô hình chấm điểm tín dụng để tiến hành xác định nhóm khách hàng tiềm năng bằng thuật toán K-means.

Mục tiêu 3: Phân nhóm khách hàng vào nhóm tiềm năng và không tiềm năng dựa trên điểm tín dụng. Từ đó công ty có thể đưa ra các chiến lược tiếp cận phù hợp với từng nhóm.

Phương pháp 3:

B1: Thu thập dữ liệu

Thu thập dữ liệu điểm tín dụng khách hàng từ mô hình chấm điểm tín dụng của Công ty TNHH MTV Viễn thông Quốc tế FPT:

- Dữ liệu này có thể bao gồm: Điểm tín dụng, thông tin cá nhân cơ bản và dữ liệu tài chính khác
- Chọn lọc các thông tin cần thiết từ dữ liệu để thực hiện việc phân loại nhóm khách hàng tiềm năng bằng thuật toán K-means

B2: Tiền xử lý dữ liệu

Các bước thực hiện:

- Làm sạch dữ liệu: Sửa hoặc xóa dữ liệu trùng lặp và những dữ liệu không liên quan có thể ảnh hưởng đến kết quả của thuật toán K-means.
- Xử lý dữ liệu bị thiếu: Tùy theo hiểu biết về bộ dữ liệu và tỉ trọng của lượng dữ liệu bị thiếu so với tổng thể, có thể loại bỏ hoặc thay thế dữ liệu bị thiếu để đảm bảo độ chính xác của dữ liệu. Sử dụng các kỹ thuật như điền giá trị trung bình, median, hoặc sử dụng mô hình dự đoán.

- Tiêu chuẩn hóa dữ liệu: Đảm bảo rằng các biến đầu vào cụ thể ở đây là điểm tin dụng có cùng phạm vi giá trị để thuật toán K-means hoạt động hiệu quả hơn. Sử dụng các kỹ thuật tiêu chuẩn hóa như Min-Max Scaling hoặc Standard Scaling.
- Biến đổi dữ liệu: Thay đổi hình thức, cấu trúc hoặc giá trị của dữ liệu nhằm tổ chức dữ liệu một cách trực quan, hiệu quả hơn có thể dùng các phương pháp Chuẩn hóa Z-score, chuyển đổi logarit,... để làm cho thuật toán K-means hoạt động tốt hơn.
- Chuẩn bị dữ liệu cho phân cụm: Sử dụng các phương pháp mã hóa như mã hóa one-hot hoặc mã hóa số, và loại bỏ các biến không cần thiết trước khi áp dụng thuật toán K-means.

B3: Phân cụm với thuật toán K-means

Thuật toán phân cụm K-means được giới thiệu năm 1957 bởi Lloyd K-means và là phương pháp phổ biến nhất cho việc phân cụm, dựa trên việc phân vùng dữ liệu.

Biểu diễn dữ liệu $D = \{x_1, x_2, \dots, x_r\}$ với x_i là vector n chiều trong không gian Euclidean. K-means phân cụm D thành K cụm dữ liệu:

- Mỗi cụm dữ liệu có một điểm trung tâm gọi là centroid.
 - K là một hằng số cho trước.
- Xác định số lượng cụm tối ưu giúp phân chia dữ liệu một cách hiệu quả và phản ánh đầy đủ các nhóm khách hàng.
 - Sử dụng chỉ số Silhouette: Đối với mỗi số lượng cụm, tính toán chỉ số Silhouette để đo độ tách biệt giữa các cụm. Chọn số lượng cụm tạo ra chỉ số Silhouette cao nhất.
 - Quá trình phân cụm khách hàng:
 - Bước 1: Khởi tạo các trung tâm ban đầu: Chọn ngẫu nhiên K điểm trong tập dữ liệu ban đầu làm các điểm trung tâm ban đầu. Đây là các điểm đại diện cho K nhóm dữ liệu

- Bước 2: Gán mỗi khách hàng vào cụm gần nhất: Với mỗi điểm dữ liệu, tính khoảng cách Euclid từ điểm đó tới K điểm trung tâm và gán điểm dữ liệu vào nhóm có điểm trung tâm gần nhất. Kết quả là một phân loại ban đầu của dữ liệu thành các nhóm.
- Bước 3: Cập nhật các trung tâm cụm: Tính toán lại vị trí của các trung tâm cụm bằng cách lấy trung bình của tất cả các điểm thuộc cùng một cụm. Cập nhật này được thực hiện sau mỗi lần phân loại.
- Bước 4: Lặp lại bước 2 và 3: Tiếp tục lặp lại quá trình gán cụm và cập nhật trung tâm cho đến khi không có sự thay đổi đáng kể trong vị trí của các trung tâm cụm hoặc đạt đến số lần lặp tối đa được định trước.
- Bước 5: Đầu ra: Kết quả của thuật toán là K điểm trung tâm cuối cùng và phân loại của từng điểm dữ liệu dựa trên vị trí của các điểm trung tâm. Các điểm trung tâm đại diện cho các nhóm trong dữ liệu.

B4: Trực quan hóa kết quả

-Sau khi áp dụng thuật toán K-means với số lượng cụm được lựa chọn có thể phân khách hàng thành các nhóm sau:

- Khách hàng VIP: Thu nhập với điểm tín dụng cao. Chi tiêu nhiều nhất và thường xuyên, sẵn sàng thúc đẩy cho công ty.
- Khách hàng tiềm năng: Điểm tín dụng cao nhưng tần suất mua hàng trung bình
- Khách hàng kỳ vọng: Điểm tín dụng trung bình nhưng chi tiêu nhiều
- Khách hàng rủi ro: Điểm tín dụng thấp

Kết quả 3:

- Sau khi phân cụm, xác định được nhóm khách hàng tiềm năng và không tiềm năng
- Phân tích các đặc trưng và hành vi của nhóm khách hàng tiềm năng để hiểu rõ hơn về họ và phát triển các chiến lược tiếp thị hoặc kinh doanh phù hợp.

Công cụ sử dụng:

- Ngôn ngữ lập trình: Python
- Thư viện: NumPy, Pandas, SkLearn, Scipy, YellowBrick

Nội dung 4: Triển khai hệ thống

Mục tiêu 4:

- Hệ thống cần có khả năng thu thập dữ liệu đầy đủ và chính xác từ nhiều nguồn khác nhau
- Hệ thống phải tích hợp một mô hình chấm điểm tín dụng đã được xác minh để đánh giá khả năng tín dụng của khách hàng, đảm bảo tính chính xác và đáng tin cậy của điểm số.
- Hệ thống cần có khả năng mở rộng để xử lý lượng lớn dữ liệu và có thể điều chỉnh để phù hợp với các yêu cầu thay đổi.
- Cung cấp các công cụ báo cáo và phân tích để theo dõi hiệu suất của các nhóm khách hàng và đánh giá để thay đổi chiến lược phù hợp.

Phương pháp 4:

- Trước hết cần có dữ liệu mẫu


```
•[1]: import pandas as pd

# Tạo dataframe từ dữ liệu mẫu
data = {
    'age': [25, 45, 35, 50, 23, 43, 36, 52],
    'salary': [5600, 8200, 9900, 13200, 8600, 18300, 23100, 11200],
    'credit_score': [450, 500, 610, 580, 460, 570, 390, 730]
}

df = pd.DataFrame(data)
df
```

[1]:

	age	salary	credit_score
0	25	5600	450
1	45	8200	500
2	35	9900	610
3	50	13200	580
4	23	8600	460
5	43	18300	570
6	36	23100	390
7	52	11200	730

	age	salary	credit_score
0	25	5600	450
1	45	8200	500
2	35	9900	610
3	50	13200	580
4	23	8600	460
5	43	18300	570
6	36	23100	390
7	52	11200	730

- Các thông tin trong dataframe trên bao gồm số tuổi (age), thu nhập cá nhân (salary) và điểm tín dụng (credit_score)

- Tiếp theo cần chuẩn hóa dữ liệu thành dataframe để có thể xử lý dữ liệu dễ dàng hơn

```
[2]: from sklearn.preprocessing import StandardScaler

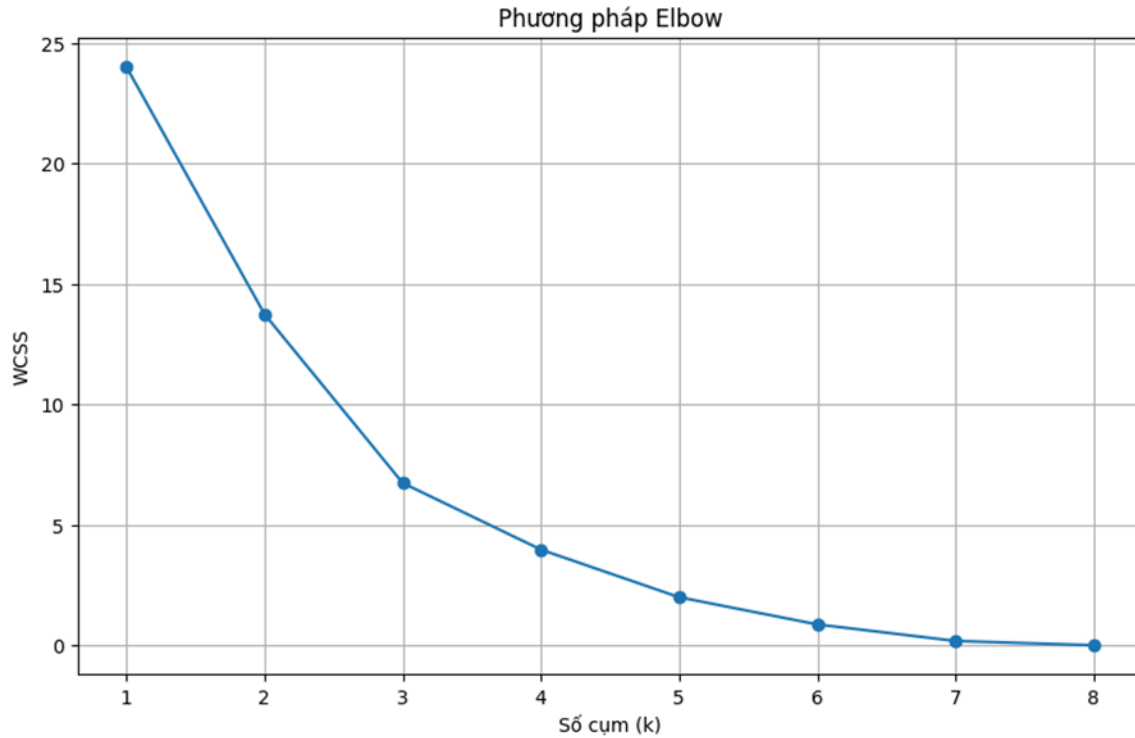
# Chuẩn hóa dữ liệu
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df)

# Chuyển dữ liệu chuẩn hóa thành DataFrame để dễ dàng xử lý hơn
df_scaled = pd.DataFrame(df_scaled, columns=df.columns)
df_scaled
```

```
[2]:
```

	age	salary	credit_score
0	-1.347529	-1.227189	-0.851982
1	0.630495	-0.748286	-0.358079
2	-0.358517	-0.435157	0.728506
3	1.125001	0.172681	0.432165
4	-1.545331	-0.674609	-0.753201
5	0.432693	1.112068	0.333384
6	-0.259616	1.996197	-1.444665
7	1.322803	-0.195706	1.913872

- Sau khi dữ liệu đã được chuẩn hóa thành công. Bước tiếp theo là áp dụng phương pháp Elbow để xác định số lượng cụm tối ưu bằng cách tính toán “tổng bình phương khoảng cách trong cụm” (WCSS) cho một số giá trị khác nhau của k từ 1 đến 8 và vẽ biểu đồ để quan sát điểm "Elbow"



- Biểu đồ Elbow trên cho thấy “tổng bình phương khoảng cách trong cụm” (WCSS) giảm đáng kể khi số lượng cụm tăng từ 1 đến 3 và sau đó bắt đầu ổn định. Do đó điểm "elbow" nằm ở khoảng $k = 3$, nơi sự giảm WCSS không còn đáng kể nên 3 là số lượng cụm tối ưu cho bộ dữ liệu này.

- Sau đó áp dụng thuật toán K-means với $k = 3$ để phân cụm dữ liệu và phân tích các cụm tìm được.

```
[6]: # Áp dụng K-means với k=3
kmeans = KMeans(n_clusters=3, random_state=42)
df['cluster'] = kmeans.fit_predict(df_scaled)

# Xem kết quả phân cụm
df.groupby('cluster').mean()
```

```
[6]:
```

	age	salary	credit_score
cluster			
0	45.0	12160.0	598.0
1	36.0	23100.0	390.0
2	24.0	7100.0	455.0

- Kết quả phân cụm của mô hình K-means với k=3, cùng với trung bình của các thuộc tính trong mỗi cụm:

+ *Cụm 0*: Khách hàng lớn tuổi (trung bình 45 tuổi), với mức thu nhập trung bình (12,160) và điểm tín dụng khá cao (598).

+ *Cụm 1*: Khách hàng (36 tuổi), với mức thu nhập cao nhất (23,100), nhưng điểm tín dụng thấp nhất (390).

+ *Cụm 2*: Khách hàng trẻ nhất (24 tuổi), thu nhập thấp nhất (7,100), và điểm tín dụng thấp (455).

→ Giả sử cụm 1 là cụm khách hàng tiềm năng

- Sau khi huấn luyện xong 2 mô hình ở trên là K-Means và Scaler thì lưu lại mô hình

- Sử dụng Flask để tạo API, xử lý yêu cầu POST, nhận dữ liệu khách hàng, áp dụng các mô hình đã huấn luyện và trả về phản hồi liệu khách hàng có phải là tiềm năng không.

```
from flask import Flask, request, jsonify
import joblib
import numpy as np

app = Flask(__name__)

# Tải các mô hình đã lưu
kmeans = joblib.load('kmeans_model.pkl')
scaler = joblib.load('scaler.pkl')

@app.route('/predict', methods=['POST'])
def predict():
    # Nhận dữ liệu từ request
    data = request.get_json()
    age = data['age']
    salary = data['salary']
    credit_score = data['credit_score']

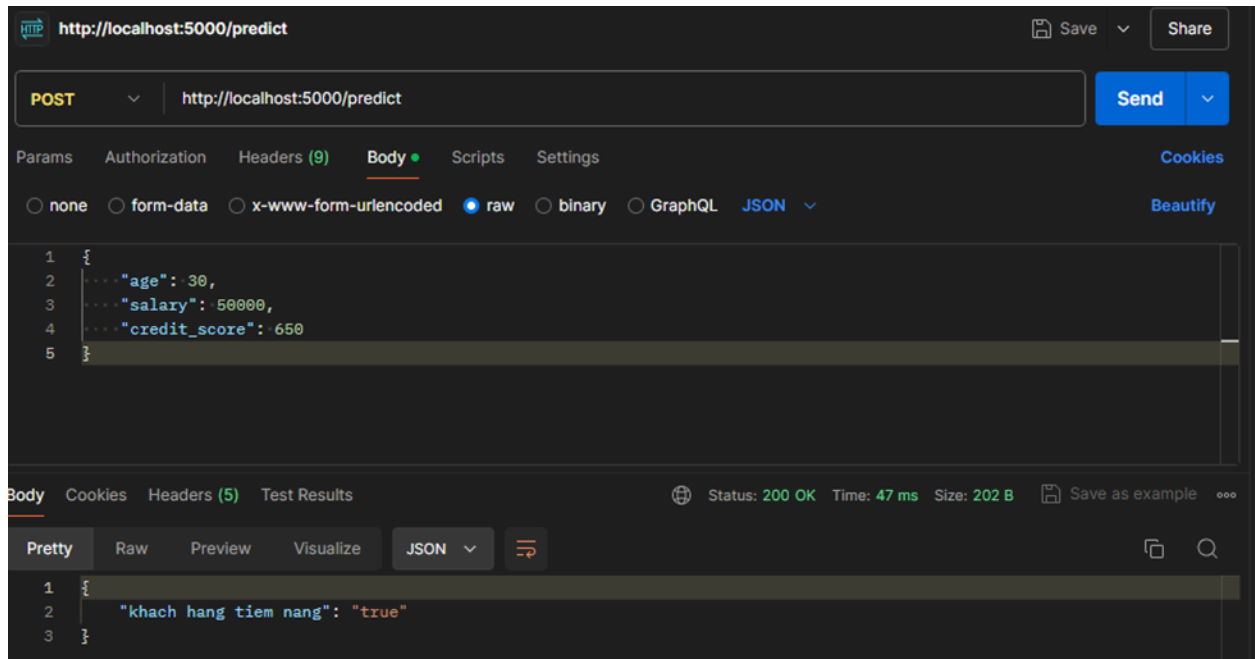
    # Chuẩn hóa dữ liệu và áp dụng mô hình
    features = np.array([[age, salary, credit_score]])
    features_scaled = scaler.transform(features)
    cluster = kmeans.predict(features_scaled)

    # Xác định khách hàng tiềm năng
    potential_customer = (cluster[0] == 1) # Giả sử cụm 1 là cụm khách hàng tiềm năng

    return jsonify({'khach hang tiem nang': str(potential_customer).lower()})

if __name__ == '__main__':
    app.run(debug=True, port=5000)
```

- Sau khi triển khai API ở trên thì gửi yêu cầu đến API với một payload JSON chứa các trường dữ liệu cần thiết như tuổi (age), thu nhập (salary) và điểm tín dụng (credit_score)



- API sẽ trả về một phản hồi cho biết liệu khách hàng có phải là khách hàng tiềm năng không

Kết quả 4:

- Hiệu suất cao: Sử dụng công nghệ tiên tiến trong cả phần cứng và phần mềm gồm phần cứng chuyên dụng hoặc dịch vụ điện toán đám mây mạnh mẽ. Bên cạnh đó sử dụng các khung tính toán hàng đầu như TensorFlow hoặc PyTorch để đảm bảo hiệu suất tối đa cho mô hình học máy.
- Dễ sử dụng: Giao diện người dùng được thiết kế thân thiện và trực quan, cho phép người dùng dễ dàng tương tác và trải nghiệm người dùng thuận lợi thông qua giao diện web hoặc mobile.
- Độ chính xác cao: Hệ thống được thử nghiệm với dữ liệu thực tế để đảm bảo độ chính xác của mô hình dự đoán. Việc thu thập phản hồi từ người dùng và điều chỉnh cần thiết giúp cải thiện độ chính xác của hệ thống.
- An toàn và bảo mật: Tuân thủ các quy định về quyền riêng tư dữ liệu là ưu tiên hàng đầu, đảm bảo rằng thông tin cá nhân của người dùng được bảo vệ và xử lý an toàn.

B3. Sản phẩm đề tài

Một ứng dụng BA Chấm điểm tín dụng dành cho Bộ phận Marketing và bán hàng

B4. Kết quả dự kiến

- Mô hình sẽ giúp phân loại khách hàng vào các nhóm dựa trên các đặc điểm như tuổi, thu nhập và điểm tín dụng. Mỗi cụm sẽ đại diện cho một phân khúc khách hàng với nhu cầu và hành vi tài chính khác nhau.
- Từ đó, cụm được xác định là cụm khách hàng tiềm năng sẽ bao gồm những khách hàng có điểm tín dụng cao
- Thông tin từ việc phân tích cụm sẽ giúp nhận diện các điểm mạnh và điểm yếu trong sản phẩm và từ đó đề xuất cải tiến.

B5. Đánh giá tính khả thi

- Ưu điểm: Dựa trên dữ liệu từ mô hình này, công ty có thể xác định và tiếp cận khách hàng tiềm năng một cách hiệu quả hơn, từ đó tạo ra mối quan hệ tốt hơn với họ. Ứng dụng sẽ giúp công ty tối ưu hóa các chiến lược tiếp thị và bán hàng để phù hợp với từng nhóm khách hàng. Điều này giúp tăng cơ hội tương tác với khách hàng đồng thời có thể tăng hiệu suất kinh doanh, lợi nhuận và tăng cường sự cạnh tranh trên thị trường. Ngoài ra công ty có thể dự báo doanh số bán hàng trong tương lai giúp quản lý dự trữ hàng hóa và tài nguyên một cách hiệu quả hơn.
- Nhược điểm: Mô hình này thường đòi hỏi một lượng lớn dữ liệu để đào tạo mạng thần kinh và để phân cụm dữ liệu bằng K-Means, điều này có thể là một thách thức nếu dữ liệu không đủ hoặc không chất lượng. Khi kết hợp nhiều phương pháp, việc đảm bảo tính nhất quán và độ tin cậy của dự đoán trở nên phức tạp hơn. Sự không nhất quán có thể xảy ra nếu các phương pháp không được tích hợp chính xác. Mô hình này cũng có nhiều siêu tham số cần được điều chỉnh và tinh chỉnh. Việc lựa chọn siêu tham số không tốt có thể dẫn đến hiệu suất kém của mô hình.
- Tiềm năng phát triển: Mở rộng áp dụng cho nhiều lĩnh vực khác như bán lẻ, ngân hàng, bảo hiểm, v.v. Giải pháp này không chỉ giới hạn cho công ty viễn thông mà

có thể áp dụng cho nhiều ngành nghề kinh doanh khác để đánh giá và tìm kiếm khách hàng tiềm năng.

C. KẾT LUẬN

Hiện nay, việc áp dụng máy học vào lĩnh vực chấm điểm tín dụng đã mở ra những cơ hội mới để nâng cao hiệu quả tìm kiếm và đánh giá khách hàng tiềm năng. Mô hình chấm điểm tín dụng được triển khai tại công ty TNHH MTV FPT không chỉ cải thiện độ chính xác trong dự đoán hành vi tài chính của khách hàng, mà còn giúp công ty có thể tối ưu hóa các chiến lược kinh doanh của mình. Công nghệ máy học với khả năng xử lý và phân tích lượng lớn dữ liệu trong thời gian thực đã thực sự thay đổi cách thức các ngân hàng và các công ty doanh nghiệp tiếp cận và đánh giá rủi ro tín dụng.

Tuy có nhiều ưu điểm nhưng mô hình này cũng cần thêm thời gian để cải thiện trong tương lai tại FPT. Kết quả của đề tài này không chỉ khẳng định vai trò của máy học trong lĩnh vực tài chính, mà còn mở ra hướng đi mới cho việc ứng dụng công nghệ thông tin trong quản lý và phát triển sản phẩm tín dụng tại FPT.

D. TÓM TẮT NỘI DUNG

STT	Nội dung thực hiện	Kết quả dự kiến	Thời gian	Kinh phí dự kiến
1	Nội dung 1	Hiểu rõ thuật toán để áp dụng các thuật toán ANNs, MARS, K-means việc xây dựng ứng dụng Chấm điểm tín dụng.	10/5-15/5	
2	Nội dung 2	Xây dựng mô hình chấm điểm tín dụng có độ chính xác cao và khả năng dự đoán tốt kết hợp thuật toán MARS và ANNs	15/5-15/7	50 triệu
3	Nội dung 3	Phân nhóm khách hàng vào nhóm tiềm năng và không tiềm năng dựa trên điểm tín dụng.bằng thuật toán K-means	15/7-30/7	50 triệu
4	Nội dung 4	Triển khai thành công hệ thống	30/7-1/9	100 triệu

		chấm điểm tín dụng để tìm kiếm khách hàng tiềm năng		
--	--	---	--	--

E. TÀI LIỆU THAM KHẢO

- [1] V. Moscato, A. Picariello, and G. Sperlí, “A benchmark of machine learning approaches for credit score prediction,” *Expert Syst. Appl.*, vol. 165, p. 113986, Mar. 2021, doi: 10.1016/j.eswa.2020.113986.
- [2] K. Tabianan, S. Velu, and V. Ravi, “K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data,” *Sustainability*, vol. 14, no. 12, Art. no. 12, Jan. 2022, doi: 10.3390/su14127243.
- [3] T.-S. Lee and I.-F. Chen, “A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines,” *Expert Syst. Appl.*, vol. 28, no. 4, pp. 743–752, May 2005, doi: 10.1016/j.eswa.2004.12.031.
- [4] “Tổng quan về Neural Network(mạng Nơ Ron nhân tạo) là gì?” Accessed: May 07, 2024. [Online]. Available: <https://itnavi.com.vn/blog/neural-network-la-gi>
- [5] “Hiểu nhanh về thuật toán K-means Clustering - Bài toán phân cụm qua ví dụ đơn giản | GMO-Z.com Vietnam Lab Center.” Accessed: May 07, 2024. [Online]. Available: <https://blog.vietnamlab.vn/untitled-11/>
- [6] “Mạng nơ ron nhân tạo: Khái niệm, cấu trúc và ứng dụng trong thực tế - Smart Factory.” Accessed: May 07, 2024. [Online]. Available: