BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

BÁO CÁO TỔNG KẾT

HỌC PHẦN: ĐỒ ÁN 2

HỆ THỐNG DỰ BÁO THÔNG MINH CHO STARTUP TRẠNG THÁI HOẠT ĐỘNG & MỨC TÀI TRỢ

Sinh viên thực hiện:

HÀ QUANG VINH	DHKL16A2HN	22174600065
LƯU NHẬT NAM	DHKL16A2HN	22174600109
KHUẤT THANH PHƯƠNG	DHKL16A2HN	22174600105

Giáo viên giảng dạy: LÊ HẰNG ANH

Hà Nội, 05/2025

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

BÁO CÁO TỔNG KẾT

HỌC PHẦN: ĐỒ ÁN 2

HỆ THỐNG DỰ BÁO THÔNG MINH CHO STARTUP TRẠNG THÁI HOẠT ĐỘNG & MỨC TÀI TRỢ

Sinh viên thực hiện:

HÀ QUANG VINH	DHKL16A2HN	22174600065
LƯU NHẬT NAM	DHKL16A2HN	22174600109
KHUẤT THANH PHƯƠNG	DHKL16A2HN	22174600105

Giáo viên giảng dạy: LÊ HẰNG ANH

Hà Nội, 05/2025

PHIẾU ĐĂNG KÝ ĐỀ TÀI

1. Tên đề tài: Hệ thống dự báo thông minh cho Startup: Trạng thái hoạt động & mức tài trợ

2. Thông tin nhóm sinh viên:

Sinh viên 1 (Nhóm trưởng):

- **Họ và tên:** Lưu Nhật Nam
- **Mã sinh viên:** 22174600109
- **Email**: lnnam.dhkl16a2hn@sv.uneti.edu.vn

Sinh viên 2:

- **Họ và tên:** Hà Quang Vinh
- **Mã sinh viên:** 22174600065
- **Email:** hqvinh.dhkl16a2hn@sv.uneti.edu.vn

Sinh viên 3:

- **Họ và tên:** Khuất Thanh Phương
- **Mã sinh viên:** 22174600105
- **Email:** ktphuong.dhkl16a2hn@sv.uneti.edu.vn

3. Tóm tắt nội dung đề tài:

Đề tài ứng dụng học máy để phân tích dữ liệu startup, gồm hai mục tiêu chính: phân loại trạng thái hoạt động của công ty (đang hoạt động, đã bị mua lại, ngừng hoạt động) và dự đoán tổng vốn đầu tư (funding_total_usd). Dữ liệu được xử lý và huấn luyện với các mô hình như Logistic Regression, Random Forest, XGBoost, SVM và KNN. Kết quả giúp đánh giá khả năng phát triển và tiềm năng gọi vốn của startup, hỗ trợ nhà đầu tư đưa ra quyết định chiến lược.

Ngày 6 tháng 4 năm 2025

Nhóm trưởng

Nam

Lưu Nhật Nam

ĐỀ CƯƠNG CHI TIẾT ĐỀ TÀI

1. Tên đề tài: Hệ thống dự báo thông minh cho Startup: Trạng thái hoạt động & mức tài trợ (Smart Forecasting System for Startups: Operational Status & Investment Levels)

2. Mục tiêu đề tài:

Mục tiêu của đề tài là xây dựng và đánh giá hai hệ thống dự đoán sử dụng các thuật toán học máy nhằm khai thác hiệu quả dữ liệu liên quan đến các công ty khởi nghiệp (startups). Thứ nhất, đề tài hướng đến việc phát triển mô hình phân loại có khả năng dự đoán trạng thái hiện tại của một công ty (ví dụ: đang hoạt động, đã bị mua lại, đã ngừng hoạt động, v.v.) dựa trên các đặc trưng như thị trường hoạt động, quốc gia, tổng vốn đầu tư, số vòng gọi vốn, và thời gian thành lập. Thứ hai, đề tài tập trung xây dựng mô hình hồi quy nhằm dự đoán tổng vốn đầu tư (funding_total_usd) mà một công ty có thể huy động, dựa trên các yếu tố đầu vào có sẵn.

Cả hai mô hình sẽ được triển khai với nhiều thuật toán học máy khác nhau như Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbors (KNN) và Support Vector Machine (SVM), từ đó đánh giá hiệu quả qua các chỉ số phù hợp (F1 Score cho phân loại, RMSE và R^2 cho hồi quy). Kết quả của đề tài kỳ vọng hỗ trợ các nhà phân tích, nhà đầu tư và cố vấn khởi nghiệp trong việc đánh giá tiềm năng phát triển của doanh nghiệp dựa trên dữ liệu định lượng.

3. Tổng quan tình hình nghiên cứu thuộc lĩnh vực đề tài:

Trong kỷ nguyên dữ liệu lớn, các công ty khởi nghiệp (startups) đã trở thành đối tượng nghiên cứu phổ biến trong các bài toán học máy nhờ vào lượng thông tin phong phú và khả năng ứng dụng cao trong thực tiễn. Việc dự đoán trạng thái hoạt động của một công ty, cũng như ước lượng tổng vốn đầu tư mà công ty đó có thể huy động, đã được nhiều nhà nghiên cứu và tổ chức đầu tư quan tâm nhằm hỗ trợ quá trình ra quyết định chiến lược và giảm thiểu rủi ro.

Các nghiên cứu trước đây chủ yếu tập trung vào việc khai thác dữ liệu từ nền tảng Crunchbase, AngelList và các nguồn dữ liệu đầu tư khác. Trong đó, các thuật toán học

máy như Logistic Regression, Decision Trees, Random Forest, Gradient Boosting (bao gồm XGBoost, LightGBM), và các mô hình deep learning đã được ứng dụng để phân loại trạng thái doanh nghiệp (operating, acquired, closed, v.v.) dựa trên các thuộc tính như thị trường mục tiêu, quốc gia, số vòng gọi vốn, loại hình đầu tư, năm thành lập, và tổng vốn đã gọi.

Ngoài ra, các bài toán hồi quy nhằm dự đoán tổng vốn đầu tư mà startup có thể huy động cũng nhận được nhiều sự quan tâm. Các mô hình như Linear Regression, Ridge, Lasso, Random Forest Regressor và XGBoost Regressor đã được sử dụng để dự đoán funding_total_usd, giúp nhà đầu tư đánh giá tiềm năng tài chính của doanh nghiệp trong giai đoạn sớm.

Bên cạnh việc xây dựng mô hình, các nghiên cứu cũng tập trung vào khâu tiền xử lý dữ liệu, bao gồm xử lý thiếu dữ liệu, mã hóa biến phân loại, chuẩn hóa dữ liệu số, và lựa chọn đặc trưng nhằm tối ưu hóa hiệu suất mô hình.

Tuy nhiên, vẫn còn nhiều thách thức trong lĩnh vực này như mất cân bằng lớp (class imbalance), nhiễu trong dữ liệu lịch sử đầu tư, và sự thay đổi nhanh chóng trong thị trường công nghệ. Do đó, việc kết hợp nhiều mô hình học máy và đánh giá hiệu suất toàn diện là cần thiết để xây dựng hệ thống dự đoán đáng tin cậy và có giá trị ứng dụng cao trong thực tế.

4. Nội dung đề tài:

Trong bối cảnh hệ sinh thái khởi nghiệp toàn cầu ngày càng phát triển mạnh mẽ, việc ra quyết định đầu tư vào một công ty startup đòi hỏi các phân tích định lượng sâu sắc và khả năng đánh giá tiềm năng tăng trưởng dựa trên dữ liệu lịch sử. Đề tài này hướng đến việc phát triển một hệ thống phân tích dựa trên học máy, phục vụ hai mục tiêu chính: phân loại trạng thái hoạt động của công ty và dự đoán tổng vốn đầu tư mà công ty có khả năng huy động được trong tương lai.

Khác với các tiếp cận truyền thống chỉ tập trung mô hình hóa dữ liệu ở mức bề mặt, đề tài này đi sâu vào phân tích cấu trúc dữ liệu, phát hiện mối quan hệ tiềm ẩn giữa các thuộc tính phi tuyến và thời gian, nhằm khai thác hiệu quả các tín hiệu quan trọng bị

ẩn dưới sự nhiễu loạn của thị trường khởi nghiệp. Trạng thái của một công ty không chỉ là kết quả của vốn đầu tư hay lĩnh vực hoạt động, mà còn chịu ảnh hưởng bởi các yếu tố khó đo lường như tính thời điểm, điều kiện địa lý, giai đoạn gọi vốn, mức độ đa dạng nguồn vốn, và cả yếu tố vùng miền. Do đó, bài toán phân loại sẽ được tiếp cận theo hướng đa chiều, với khả năng trích xuất đặc trưng nâng cao, sử dụng các kỹ thuật như feature interaction, embedding cho biến phân loại có nhiều giá trị, và phân tích mô hình theo khung thời gian.

Đối với bài toán hồi quy, thay vì chỉ ước lượng giá trị tuyệt đối của tổng vốn đầu tư (funding_total_usd), đề tài đặt trọng tâm vào khả năng mô hình hóa biến động vốn theo các nhóm startup khác nhau (theo ngành nghề, khu vực, số vòng gọi vốn). Việc này cho phép đánh giá tính chất phi tuyến trong quá trình gọi vốn – một yếu tố mà các mô hình tuyến tính đơn giản thường bỏ qua. Ngoài các mô hình chuẩn như Random Forest và XGBoost Regressor, đề tài còn xem xét các kỹ thuật như quantile regression (để dự báo biên độ kỳ vọng), ensemble stacking và kỹ thuật kiểm định chéo phân tầng nhằm tăng tính ổn định và độ khái quát của mô hình.

Một khía cạnh quan trọng khác được lồng ghép trong đề tài là đánh giá tác động của dữ liệu mất cân bằng (class imbalance) đối với mô hình phân loại. Thay vì áp dụng đơn thuần các kỹ thuật như oversampling hay undersampling, đề tài thử nghiệm các phương pháp hiện đại như SMOTE kết hợp phân cụm (Cluster-SMOTE), cũng như chiến lược gán trọng số động cho các lớp thiểu số trong quá trình huấn luyện.

Cuối cùng, từ góc độ ứng dụng, đề tài hướng đến xây dựng một khuôn mẫu có khả năng tích hợp vào hệ thống đánh giá rủi ro cho các quỹ đầu tư, vườn ươm doanh nghiệp và các nền tảng hỗ trợ startup. Đây là bước tiến từ việc "dự đoán kết quả" sang "hỗ trợ định hình quyết định", một mục tiêu cốt lõi trong khoa học dữ liệu hiện đại.

5. Phương pháp thực hiện:

Quy trình thực hiện đề tài được triển khai qua các bước chính theo quy trình khoa học dữ liệu chuẩn, từ tiền xử lý dữ liệu đến xây dựng và đánh giá mô hình học máy. Trước hết, dữ liệu thô từ nguồn Crunchbase được làm sạch và chuẩn hóa. Các giá trị thiếu được xử lý tùy theo loại biến (xóa, điền trung bình, hoặc mô hình hóa lại), biến phân loại được mã hóa bằng one-hot encoding hoặc embedding, còn các biến số được chuẩn hóa theo phân phối chuẩn hoặc log-transform để xử lý dữ liệu lệch.

Tiếp theo, phân tích khám phá dữ liệu (EDA) được thực hiện để tìm hiểu mối quan hệ giữa các thuộc tính và nhãn mục tiêu, đồng thời phát hiện các vấn đề như mất cân bằng lớp, dữ liệu ngoại lai hoặc đa cộng tuyến. Sau đó, dữ liệu được tách thành hai bộ riêng biệt cho hai bài toán: phân loại trạng thái công ty và hồi quy tổng vốn đầu tư.

Trong bước xây dựng mô hình, các thuật toán học máy hiện đại được áp dụng và so sánh:

- Phân loại: Logistic Regression, Random Forest, XGBoost, KNN và SVM.
- Hồi quy: Linear Regression, Random Forest Regressor, XGBoost Regressor, Ridge và Lasso.

Mỗi mô hình được huấn luyện với quy trình kiểm định chéo (cross-validation), đồng thời tinh chỉnh siêu tham số bằng GridSearchCV. Đánh giá mô hình dựa trên các chỉ số phù hợp: Accuracy, F1 Score cho phân loại; MAE, RMSE, R^2 cho hồi quy.

Cuối cùng, kết quả được tổng hợp và trực quan hóa bằng biểu đồ so sánh, đồng thời phân tích chuyên sâu nhằm đưa ra khuyến nghị chiến lược cho các bên liên quan.

6. Phân công công việc:

STT	Họ và tên	Mã sinh viên	Nội dung công việc được phân công
1	Lưu Nhật Nam	22174600109	Tiền xử lý dữ liệu thô (làm sạch, mã hóa, chuẩn hóa) Phân tích khám phá dữ liệu (EDA) Xây dựng & đánh giá 5 mô hình phân loại Trực quan hóa hiệu suất mô hình phân loại
2	Khuất Thanh Phương	22174600105	Xây dựng 5 mô hình hồi quy (Linear, Ridge, Lasso, RF, XGBoost) Đánh giá theo MAE, RMSE, R^2 Vẽ biểu đồ so sánh các mô hình hồi quy Rút ra kết luận về hiệu quả dự báo
3	Hà Quang Vinh	22174600065	Tổng quan đề tài, xây dựng khung nghiên cứu Viết các phần: mục tiêu đề tài, tổng quan nghiên cứu, phương pháp thực hiện, dự kiến kết quả Tổng hợp toàn bộ kết quả của nhóm Thiết kế biểu đồ, bảng biểu minh họa Đề xuất,xây dựng ứng dụng mô hình trong thực tế Chịu trách nhiệm trình bày báo cáo cuối cùng.

7. Dự kiến kết quả đạt được:

Đề tài kỳ vọng mang lại hai nhóm kết quả chính: (1) các mô hình dự đoán chính xác phục vụ bài toán phân loại và hồi quy trong lĩnh vực startup, và (2) các hiểu biết sâu sắc từ dữ liệu hỗ trợ ra quyết định chiến lược.

Với bài toán phân loại, mục tiêu là xây dựng mô hình dự đoán trạng thái hoạt động của công ty (operating, acquired, closed) với Accuracy > 80% và F1 Score > 0.75, ngay cả khi dữ liệu mất cân bằng lớp. Các mô hình như Random Forest và XGBoost được kỳ vọng sẽ vượt trội nhờ khả năng xử lý dữ liệu phi tuyến. Qua phân tích feature importance, đề tài sẽ xác định các yếu tố ảnh hưởng lớn đến sự sống còn của startup, như thị trường, quốc gia, giai đoạn gọi vốn.

Đối với bài toán hồi quy, mục tiêu là mô hình có MAE, RMSE thấp và $R^2 > 0.75$, giúp dự đoán chính xác tổng vốn đầu tư. Random Forest Regressor hoặc XGBoost Regressor được dự đoán hiệu quả hơn mô hình tuyến tính nhờ khả năng nắm bắt quan hệ phức tạp giữa các biến.

Ngoài ra, đề tài cung cấp các trực quan hóa như biểu đồ hiệu suất, ma trận nhầm lẫn, đồ thị tương quan, hỗ trợ nhà đầu tư và startup trong việc ra quyết định chiến lược.

Tóm lại, đề tài không chỉ phát triển mô hình học máy hiệu quả mà còn khai thác tri thức từ dữ liệu đầu tư, góp phần nâng cao hiệu quả trong kỷ nguyên kinh tế số.

Ngày 6 tháng 4 năm 2025

Nhóm trưởng

Nam

Lưu Nhật Nam

MỞ ĐẦU

Trong bối cảnh hệ sinh thái khởi nghiệp toàn cầu phát triển mạnh mẽ, việc dự đoán chính xác trạng thái hoạt động và mức tài trợ của các startup đã trở thành nhu cầu cấp thiết đối với các nhà đầu tư, quỹ mạo hiểm và chính các doanh nghiệp khởi nghiệp. Sự biến động mạnh mẽ của thị trường cùng tính cạnh tranh khốc liệt khiến tỷ lệ thất bại của các startup luôn ở mức cao, đồng thời làm gia tăng nhu cầu về các công cụ phân tích dữ liệu tiên tiến có khả năng đưa ra những dự báo chính xác và những hiểu biết sâu sắc.

Đề tài này tập trung phát triển một hệ thống dự báo thông minh toàn diện, kết hợp những phương pháp học máy hiện đại với kỹ thuật khai phá dữ liệu chuyên sâu, nhằm giải quyết đồng thời hai bài toán quan trọng: (1) phân loại trạng thái hoạt động của startup (đang hoạt động, được mua lại, đóng cửa, hoặc IPO) và (2) dự đoán chính xác mức tài trợ dựa trên các đặc điểm kinh doanh và lịch sử gọi vốn. Hệ thống được kỳ vọng sẽ đạt độ chính xác cao (Accuracy > 80%, F1-Score > 0.75 cho bài toán phân loại và $R^2 > 0.75$ cho bài toán hồi quy) ngay cả với những bộ dữ liệu mất cân bằng - một thách thức phổ biến trong lĩnh vực này.

Các mô hình tiên tiến như XGBoost, LightGBM và Random Forest sẽ được nghiên cứu ứng dụng nhờ khả năng xử lý hiệu quả dữ liệu phi tuyến tính và các tương tác phức tạp giữa các biến. Đặc biệt, hệ thống không chỉ dừng lại ở việc đưa ra các dự đoán mà còn tập trung vào khả năng giải thích mô hình thông qua các kỹ thuật như SHAP values và feature importance, giúp xác định rõ những yếu tố then chốt ảnh hưởng đến sự thành bại của startup như: lĩnh vực hoạt động, quốc gia, số vòng gọi vốn, kinh nghiệm đội ngũ sáng lập...

Với ý nghĩa thực tiễn cao, đề tài không chỉ góp phần nâng cao hiệu quả hoạt động của hệ sinh thái khởi nghiệp mà còn mở ra hướng nghiên cứu mới trong việc ứng dụng trí tuệ nhân tạo vào lĩnh vực tài chính - đầu tư. Kết quả nghiên cứu sẽ là nền tảng quan trọng cho các giải pháp công nghệ tiên tiến phục vụ phát triển kinh tế số trong tương lai.

Đồ án bao gồm các phần được phân chương như sau:

Chương 1: Bối cảnh và bài toán phân tích

Chương 2: Cơ sở lý thuyết

Chương 3: Thực nghiệm

Chương 4: Kết quả đạt được

Chương 5: Kết luận, ưu điểm, nhược điểm, định hướng phát triển

MỤC LỤC

MỤC LỤC HÌNH VẼ

CHƯƠNG 1. BỐI CẢNH VÀ BÀI TOÁN PHÂN TÍCH

1.1. BỐI CẢNH NGHIÊN CỨU

1.1.1. Bài toán đặt ra trong đầu tư Startup và vai trò của dữ liệu

Trong thập kỷ gần đây, hệ sinh thái khởi nghiệp đã chứng kiến sự tăng trưởng mạnh mẽ, đóng góp đáng kể vào sự đổi mới sáng tạo và phát triển kinh tế toàn cầu. Tuy nhiên, thống kê cho thấy hơn 90% startup thất bại trong vòng 5 năm đầu tiên, đặt ra bài toán nan giải cho các nhà đầu tư mạo hiểm (VCs) và quỹ đầu tư trong việc đánh giá chính xác tiềm năng và rủi ro của từng dự án.

Một trong những thách thức cốt lõi là thiếu phương pháp định lượng khách quan để dự đoán khả năng tồn tại (operational status) và năng lực huy động vốn (funding potential) của startup. Các đánh giá truyền thống chủ yếu dựa trên kinh nghiệm cá nhân hoặc phân tích định tính, dẫn đến sai lệch trong ra quyết định. Trong khi đó, sự bùng nổ của dữ liệu lớn (Big Data) và trí tuệ nhân tạo (AI) mở ra cơ hội ứng dụng các mô hình học máy (machine learning) để tối ưu hóa quy trình đầu tư.

Xuất phát từ thực tiễn này, nghiên cứu đề xuất xây dựng hệ thống dự báo thông minh dựa trên tập dữ liệu đa chiều về startup toàn cầu, bao gồm: lĩnh vực hoạt động (industry), quy mô tài trợ (funding amount), giai đoạn gọi vốn (funding round), và trạng thái hiện tại (active/acquired/closed). Kết quả đầu ra không chỉ hỗ trợ nhà đầu tư định lượng rủi ro mà còn góp phần nâng cao hiệu quả phân bổ vốn trong hệ sinh thái khởi nghiệp.

1.1.2. Vai trò của quyết định dựa trên dữ liệu (Data-Driven Decision Making) trong đầu tư Startup

Trong bối cảnh thị trường startup đặc thù với tính bất định cao, Data-Driven Decision Making (DDDM) đã trở thành yếu tố then chốt để giảm thiểu rủi ro và tối ưu lợi nhuận. Khác với phương pháp truyền thống dựa trên intuition (trực giác), DDDM cho phép nhà đầu tư:

- Định lượng hóa các chỉ số tiềm năng (vd: mối tương quan giữa số vòng gọi vốn và tỷ lệ thành công).
- Nhận diện xu hướng thông qua phân tích lịch sử dữ liệu (vd: ngành nào có tỷ lệ M&A cao nhất).
- Tự động hóa đánh giá bằng các thuật toán phân loại (classification) và hồi quy (regression).

Nghiên cứu này tập trung vào việc triển khai các kỹ thuật Advanced Analytics (phân tích nâng cao) trên dataset đa nguồn, bao gồm:

- Dự đoán trạng thái hoạt động (status prediction) sử dụng mô hình Supervised Learning.
- Ước lượng mức tài trợ tối ưu (funding estimation) dựa trên Random Forest hoặc XGBoost.

1.2. TÍNH BẤT ĐỊNH TRONG DỰ ĐOÁN THÀNH CÔNG STARTUP

Bản chất phức tạp của hệ sinh thái khởi nghiệp: Thành công của một startup phụ thuộc vào một ma trận các yếu tố động và khó định lượng, bao gồm:

- Yếu tố nội tại: Mô hình kinh doanh (business model), năng lực đội ngũ (team competency), công nghệ cốt lõi (core technology).
- Yếu tố bên ngoài: Xu hướng thị trường (market trends), cạnh tranh ngành (industry rivalry), chính sách vĩ mô (macroeconomic policies).
- Yếu tố ngẫu nhiên: Thời điểm ra mắt (timing), may mắn (luck), hoặc các sự kiện bất khả kháng (black swan events).

Sự tương tác phi tuyến (non-linear interaction) giữa các yếu tố này khiến việc dự đoán chính xác khả năng tồn tại hay phát triển của startup trở thành bài toán có độ bất định (uncertainty) cao, thậm chí với các mô hình AI tiên tiến.

1.3. PHƯƠNG PHÁP NGHIÊN CỨU

Đề tài sử dụng các phương pháp nghiên cứu sau:

- Phân tích tài liệu:
  - Nghiên cứu các tài liệu khoa học, báo cáo ngành, và dữ liệu từ Crunchbase để hiểu rõ cấu trúc dữ liệu startup và các phương pháp học máy.
- Phương pháp thực nghiệm:
  - Tiền xử lý dữ liệu: Xử lý giá trị thiếu, loại bỏ outliers, chuẩn hóa, và mã hóa biến phân loại.
  - Huấn luyện mô hình: Thử nghiệm các mô hình phân loại (Logistic Regression, Random Forest, XGBoost, LightGBM) và hồi quy (Linear Regression, Random Forest Regressor, XGBoost Regressor).
  - Đánh giá mô hình: Sử dụng các chỉ số như Accuracy, F1 Score (phân loại) và RMSE, R^2 (hồi quy); áp dụng cross-validation để đảm bảo độ tin cậy.
  - Phân tích định tính: Phân tích feature importance để xác định các yếu tố quan trọng ảnh hưởng đến trạng thái và vốn đầu tư.

- Phương pháp so sánh: So sánh hiệu suất các mô hình để chọn mô hình tối ưu cho từng bài toán.

1.4. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

- Đối tượng nghiên cứu:
  - Dữ liệu về 54,294 startup từ tập dữ liệu Investments trên Kaggle, bao gồm thông tin về trạng thái hoạt động, tổng vốn đầu tư, thị trường, quốc gia, năm thành lập, và các loại hình gọi vốn.
  - Các mô hình học máy và kỹ thuật tiền xử lý dữ liệu để dự đoán trạng thái hoạt động và mức vốn đầu tư.

- Phạm vi nghiên cứu:
  - Không gian: Các startup trên toàn cầu, chủ yếu từ các quốc gia phát triển như Mỹ, Anh, Canada, Đức, và Israel, với dữ liệu từ năm 1995 trở đi.
  - Thời gian: Dữ liệu gọi vốn và trạng thái hoạt động từ năm 1921 đến 2015, tập trung vào các startup thành lập sau năm 1995 để đảm bảo tính phù hợp với bối cảnh công nghệ hiện đại.
  - Nội dung: Tập trung vào hai bài toán chính: phân loại trạng thái (operating, acquired, closed) và dự đoán tổng vốn đầu tư (funding_total_usd), với trọng tâm là cải thiện độ chính xác và khả năng ứng dụng thực tiễn.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. TỔNG QUAN VỀ HỆ SINH THÁI STARTUP

Hệ sinh thái khởi nghiệp là một môi trường năng động, nơi các startup trải qua các giai đoạn phát triển khác nhau, từ hình thành ý tưởng đến trưởng thành hoặc thoái vốn. Hiểu rõ các giai đoạn này và các yếu tố ảnh hưởng đến thành công hay thất bại của startup là nền tảng để xây dựng hệ thống dự báo thông minh dựa trên dữ liệu.

2.1.1. Các giai đoạn phát triển startup và vòng gọi vốn

Startup tiến hóa qua các giai đoạn phát triển, mỗi giai đoạn gắn liền với các nhu cầu tài chính và vòng gọi vốn riêng biệt.

- Ở giai đoạn ý tưởng (Pre-Seed), các nhà sáng lập tập trung xây dựng sản phẩm tối thiểu (MVP) để thử nghiệm ý tưởng, thường sử dụng vốn tự có hoặc từ bạn bè, gia đình, và nhà đầu tư thiên thần, với quy mô đầu tư thường dưới 500,000 USD.
- Khi bước vào giai đoạn khởi đầu (Seed), startup hoàn thiện sản phẩm, kiểm chứng thị trường, và xây dựng đội ngũ, thu hút vốn từ quỹ đầu tư mạo hiểm hoặc các chương trình ươm mầm với mức đầu tư từ 500,000 đến 2 triệu USD. Giai đoạn tăng trưởng bao gồm các vòng Series A, B, và C, trong đó Series A (2-15 triệu USD) tập trung mở rộng thị trường và cải thiện sản phẩm, Series B (10-50 triệu USD) đẩy mạnh quy mô hoạt động và thương hiệu, còn Series C trở lên (từ 50 triệu USD) hướng đến mở rộng quốc tế, sáp nhập, hoặc chuẩn bị phát hành cổ phiếu (IPO).
- Cuối cùng, ở giai đoạn trưởng thành, startup có thể đạt trạng thái bị mua lại (acquired), IPO, hoặc ngừng hoạt động (closed). Từ góc độ khoa học dữ liệu, dữ liệu về các vòng gọi vốn, chẳng hạn như seed, venture, round_A, round_B, round_C, là những đặc trưng quan trọng phản ánh mức độ trưởng thành và tiềm năng tài chính của startup.

Tuy nhiên, dữ liệu này thường không đầy đủ hoặc thiếu đồng nhất, đòi hỏi các kỹ thuật tiền xử lý như điền giá trị thiếu, chuẩn hóa định dạng, và xử lý giá trị ngoại lai để đảm bảo chất lượng đầu vào cho mô hình học máy.

2.1.2. Yếu tố ảnh hưởng đến thành công/thất bại của startup

Thành công hay thất bại của một startup phụ thuộc vào sự kết hợp phức tạp của nhiều yếu tố, từ đặc điểm nội tại của công ty đến các lực lượng bên ngoài và các sự kiện ngẫu nhiên.

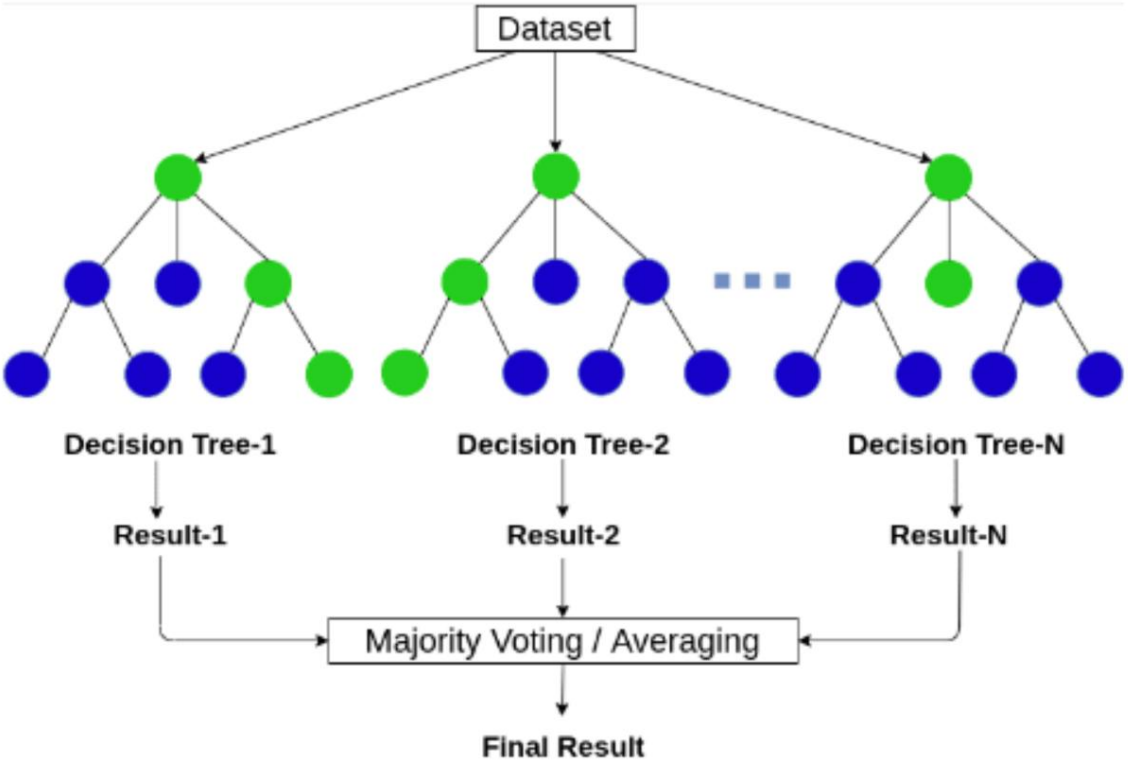
- Về mặt nội tại, mô hình kinh doanh đóng vai trò cốt lõi, với tính bền vững, khả năng mở rộng, và sự phù hợp với nhu cầu thị trường (như các mô hình SaaS hoặc marketplace) là yếu tố quyết định.
- Năng lực đội ngũ, bao gồm kinh nghiệm của nhà sáng lập, kỹ năng quản lý, và khả năng đổi mới, là động lực thúc đẩy sự phát triển.
- Công nghệ cốt lõi, với mức độ tiên tiến, khả năng bảo vệ bằng sáng chế, và tính cạnh tranh, giúp startup tạo ra lợi thế trên thị trường.
- Bên ngoài, xu hướng thị trường, chẳng hạn như tốc độ tăng trưởng của ngành (ví dụ, AI tăng 20% mỗi năm), nhu cầu khách hàng, và chu kỳ kinh tế, định hình cơ hội phát triển.
- Cạnh tranh ngành, bao gồm số lượng đối thủ, rào cản gia nhập, và mức độ khác biệt hóa, ảnh hưởng trực tiếp đến khả năng tồn tại của startup.
- Các chính sách vĩ mô, như lãi suất, ưu đãi thuế, hoặc quy định khởi nghiệp, cũng đóng vai trò quan trọng trong việc hỗ trợ hoặc kìm hãm sự phát triển. Ngoài ra, các yếu tố ngẫu nhiên, như thời điểm ra mắt sản phẩm (ví dụ, Zoom tận dụng đại dịch), may mắn trong việc gặp nhà đầu tư phù hợp, hoặc các sự kiện bất khả kháng như khủng hoảng kinh tế và thiên tai, có thể thay đổi hoàn toàn quỹ đạo của startup.

Từ góc độ khoa học dữ liệu, những yếu tố này đặt ra các thách thức lớn trong mô hình hóa. Dữ liệu định lượng, như số vòng gọi vốn, thị trường, hoặc quốc gia, dễ thu thập nhưng không thể hiện đầy đủ các yếu tố định tính như năng lực đội ngũ hay mức độ sáng tạo. Sự tương tác phi tuyến giữa các yếu tố yêu cầu sử dụng các mô hình học máy tiên tiến, chẳng hạn như XGBoost hoặc mạng nơ-ron, để nắm bắt các mối quan hệ phức tạp. Hơn nữa, tính bất định cao của hệ sinh thái khởi nghiệp đòi hỏi các kỹ thuật như Probabilistic Models hoặc Monte Carlo Simulation để mô hình hóa rủi ro và cung cấp dự đoán kèm theo độ tin cậy. Cơ sở lý thuyết này định hướng việc thiết kế hệ thống dự báo, nhấn mạnh vào việc tích hợp các đặc trưng đa dạng và xử lý các thách thức dữ liệu một cách hiệu quả.

2.2. THUẬT TOÁN RANDOM FOREST

2.2.1. Giới thiệu về Random Forest

Random Forest là một thuật toán học máy thuộc nhóm học có giám sát (supervised learning) và được sử dụng phổ biến trong các bài toán phân loại (classification) và hồi quy (regression). Thuật toán này là một dạng của tập hợp học (ensemble learning), nơi mà nhiều mô hình yếu (weak learners), cụ thể là các cây quyết định (decision trees), được kết hợp lại để tạo thành một mô hình mạnh mẽ hơn.



Hình 2.1: Mô hình Rừng Ngẫu nhiên (Random Forest) trong Học Máy

Tại sao lại là "Random"?

Thuật ngữ "Random" trong Random Forest xuất phát từ hai yếu tố chính:

Ngẫu nhiên trong chọn mẫu

Thay vì sử dụng toàn bộ dữ liệu huấn luyện để xây dựng từng cây quyết định, thuật toán Random Forest chọn một mẫu ngẫu nhiên từ tập dữ liệu (với hoàn lại) để xây dựng mỗi cây. Kỹ thuật này được gọi là Bagging (Bootstrap Aggregating). Bagging giúp giảm thiểu phương sai của mô hình, cải thiện độ chính xác tổng thể.

Ngẫu nhiên trong chọn đặc trưng

Khi tạo các nút trong mỗi cây, chỉ một tập con ngẫu nhiên của tất cả các đặc trưng được xem xét để chọn đặc trưng tốt nhất tại mỗi bước. Điều này giúp các cây quyết định đa dạng hơn, giảm thiểu hiện tượng overfitting và đảm bảo rằng các cây không bị phụ thuộc quá mức vào một đặc trưng cụ thể nào đó.

2.2.2. Cơ chế hoạt động của Random Forest

2.2.2.1. Giới thiệu

Random Forest bao gồm nhiều cây quyết định (Decision Trees). Mỗi cây là một mô hình dự đoán độc lập và đưa ra một dự đoán. Đối với bài toán phân loại, Random Forest sẽ lấy kết quả dự đoán của từng cây và chọn kết quả nào xuất hiện nhiều nhất (majority vote). Đối với bài toán hồi quy, kết quả cuối cùng là giá trị trung bình của các dự đoán từ tất cả các cây.

- Ví dụ, giả sử chúng ta có 100 cây quyết định. Đối với một mẫu mới, nếu 60 cây dự đoán rằng mẫu đó thuộc lớp A, và 40 cây dự đoán rằng mẫu đó thuộc lớp B, thì Random Forest sẽ dự đoán rằng mẫu đó thuộc lớp A (vì nó nhận được số phiếu cao hơn).

2.2.2.2. Công thức tổng quát

Một cây quyết định trong Random Forest thực hiện phân loại hoặc hồi quy bằng cách chia nhỏ không gian đặc trưng thành các vùng con. Các phân vùng này được xác định dựa trên các điều kiện phân tách tại mỗi nút trong cây. Giả sử có một đặc trưng X và một ngưỡng phân tách t, việc phân tách tại một nút có thể được biểu diễn bằng cách chọn một hàm chỉ thị I, trong đó:

$$I(X \leq t) \text{ và } I(X > t) \qquad (1)$$

Nếu đặc trưng X tại mẫu đó nhỏ hơn hoặc bằng t, mẫu sẽ được chuyển đến nhánh trái; ngược lại, nó sẽ được chuyển đến nhánh phải. Quá trình này tiếp tục cho đến khi đạt đến một nút lá, nơi giá trị đầu ra của nút đó được sử dụng làm dự đoán.

2.2.2.3. Hàm mất mát và độ không thuần khiết

Quá trình xây dựng cây quyết định trong Random Forest liên quan đến việc tối ưu hóa một hàm mất mát, thường là giảm thiểu độ không thuần khiết (impurity) của các nút. Đối với bài toán phân loại, độ không thuần khiết thường được đo bằng chỉ số Gini hoặc entropy.

Chỉ số Gini là một cách đo lường độ không thuần khiết của một nút. Công thức tính chỉ số Gini cho một nút t là:

$$\text{Gini(t)} = 1 - \sum_{i=1}^{C} p_i^2$$

Trong đó, pi là xác suất của việc một mẫu thuộc lớp i tại nút t, và C là tổng số lớp. Một nút thuần khiết (tức là tất cả các mẫu đều thuộc một lớp) sẽ có chỉ số Gini bằng 0.

Entropy là một thước đo khác về độ không thuần khiết, và nó được sử dụng trong việc xây dựng cây quyết định theo phương pháp ID3 hoặc C4.5. Công thức tính entropy tại một nút t là:

$$\text{Entropy(t)} = -\sum_{i=1}^{C} p_i log_2(p_i)$$

Tương tự như chỉ số Gini, entropy đạt giá trị nhỏ nhất khi nút hoàn toàn thuần khiết.

Giảm độ không thuần khiết (Information Gain)

Khi một đặc trưng được chọn để phân tách tại một nút, mục tiêu là làm giảm độ không thuần khiết của các nút con so với nút cha. Sự giảm này, được gọi là Information Gain (đối với entropy) hoặc Gini Gain (đối với chỉ số Gini), được tính như sau:

$$Information\ Gain = Impurity(t) - \sum_{k \in \{left, right\}} \frac{N_k}{N} Impurity(t_k)$$

Trong đó:

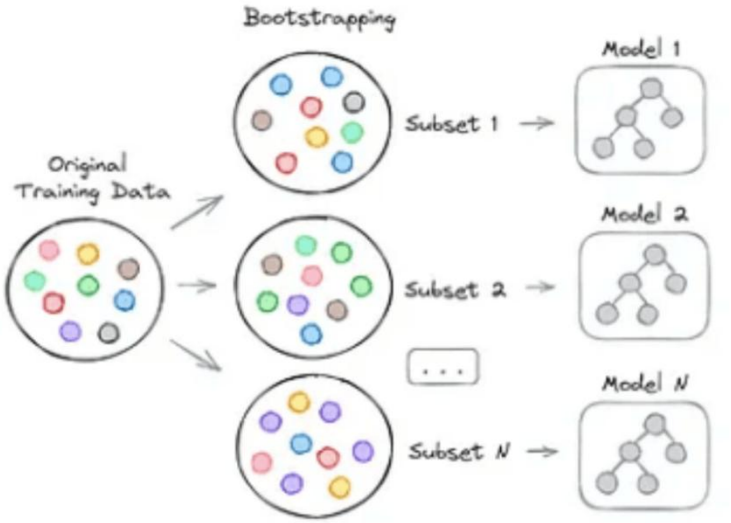
- Impurity(t) là độ không thuần khiết tại nút cha.
- Impurity(tk) là độ không thuần khiết tại các nút con sau khi phân tách.
- N là số lượng mẫu tại nút cha, và Nk là số lượng mẫu tại nút con k.

2.2.2.4. Bagging và ngẫu nhiên hóa

Trong Random Forest, hai khía cạnh quan trọng của tính ngẫu nhiên giúp tăng cường khả năng tổng quát hóa của mô hình:

Thuật toán Random Forest sử dụng phương pháp Bagging (Bootstrap Aggregating) để xây dựng mỗi cây quyết định. Thay vì sử dụng toàn bộ tập dữ liệu huấn luyện, mỗi cây

được huấn luyện trên một mẫu ngẫu nhiên từ tập dữ liệu, với việc lấy mẫu có hoàn lại (tức là một mẫu có thể được chọn nhiều lần).



Hình 2.2: Nguyên lí hoạt động của Bagging (Bootstrap Aggregating)

- Ngẫu nhiên hóa đặc trưng

Tại mỗi bước chia tách trong cây, một tập con ngẫu nhiên của các đặc trưng được xem xét để tìm đặc trưng tốt nhất. Điều này làm cho mỗi cây khác biệt hơn và giảm thiểu sự phụ thuộc vào một số đặc trưng cụ thể, từ đó giảm nguy cơ overfitting.

2.2.2.5. Đánh giá mức độ quan trọng của đặc trưng

Trong Random Forest, mức độ quan trọng của các đặc trưng được đánh giá dựa trên mức giảm độ không thuần khiết (impurity) mà đặc trưng đó đóng góp khi được chọn làm đặc trưng phân tách. Đối với mỗi cây, tổng mức giảm impurity trên toàn bộ cây được tính cho mỗi đặc trưng, và sau đó được trung bình hóa qua tất cả các cây trong rừng.

Một cách khác để đánh giá tầm quan trọng của đặc trưng là sử dụng phương pháp *Permuted Feature Importance*, trong đó các giá trị của một đặc trưng cụ thể được xáo trộn ngẫu nhiên và mức giảm trong độ chính xác của mô hình được sử dụng để đánh giá mức độ quan trọng của đặc trưng đó.

Tóm lại, mức độ quan trọng của một đặc trưng Xj có thể được tính bằng công thức:

$$Feature\ Importance(X_j) = \frac{1}{B}\sum_{b=1}^{B}\sum_{t \in nodes}\Delta I(t)$$

Trong đó:

- ΔI(t) là mức giảm impurity tại nút t khi sử dụng đặc trưng Xj.
- B là tổng số cây trong rừng.
- nodes là tập hợp các nút trong cây b nơi Xj được sử dụng.

2.3. THUẬT TOÁN XGBOOST

XGBoost là một thuật toán học máy thuộc về thể loại học tập tổng hợp, cụ thể là khuôn khổ tăng cường độ dốc. Nó sử dụng cây quyết định làm trình học cơ sở và sử dụng các kỹ thuật chính quy hóa để tăng cường khái quát hóa mô hình. XGBoost nổi tiếng với hiệu quả tính toán, cung cấp khả năng xử lý hiệu quả, phân tích tầm quan trọng của tính năng sâu sắc và xử lý liền mạch các giá trị bị thiếu. Đây là thuật toán phù hợp cho nhiều tác vụ, bao gồm hồi quy, phân loại và xếp hạng.

2.3.1. XGBOOST trong Machine Learning

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.

Hình 2.3: Mô tả quá trình học của XGBoost

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.

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.

2.3.2. Một số cơ chế của XGBoost

XGBoost đã mang đến những nâng cấp giá trị so với phương pháp gradient boosting cơ bản thông qua một số cơ chế đổi mới:

- **Điều chuẩn hóa (Regularization):** Ngay trong quá trình huấn luyện mô hình, XGBoost đã tích hợp hai dạng điều chuẩn (L1 và L2) vào hàm mục tiêu. Việc này tạo ra một "rào cản" chống lại sự phức tạp quá mức của mô hình, qua đó góp phần đáng kể vào việc hạn chế tình trạng quá khớp, đặc biệt phát huy tác dụng khi làm việc với những bộ dữ liệu có kích thước lớn hoặc bản chất phức tạp.

- **Tính toán song song (Parallelization):** Một ưu điểm vượt trội so với nhiều phiên bản gradient boosting trước đó là XGBoost được thiết kế để tận dụng khả năng xử lý song song của phần cứng. Điều này cho phép thuật toán thực hiện nhiều phép tính đồng thời, giúp rút ngắn đáng kể thời gian cần thiết cho việc huấn luyện mô hình.

- **Tối ưu bộ nhớ (Memory Efficiency):** XGBoost triển khai các cấu trúc dữ liệu và thuật toán được tối ưu hóa để sử dụng bộ nhớ một cách hiệu quả. Nhờ vậy, nó có khả năng xử lý các tập dữ liệu rất lớn mà không đòi hỏi quá nhiều tài nguyên bộ nhớ, làm tăng tính thực tiễn khi áp dụng vào các bài toán quy mô lớn.

- **Xử lý giá trị khuyết (Missing Values Handling):** Thuật toán này được trang bị khả năng xử lý tự động các giá trị bị thiếu trong dữ liệu đầu vào. Người dùng không nhất thiết phải thực hiện các bước tiền xử lý phức tạp để lấp đầy hoặc loại bỏ các giá trị này, giúp tiết kiệm thời gian và công sức trong giai đoạn chuẩn bị dữ liệu.

2.3.3. Cấu trúc của XGBoost

Để nắm bắt cách thức XGBoost vận hành một cách thấu đáo, việc xem xét kỹ lưỡng cấu trúc nền tảng cũng như các yếu tố cấu thành chủ chốt của thuật toán này là điều cần thiết.

2.3.3.1. Decision Trees (Cây quyết định)

Trong kiến trúc của XGBoost, cây quyết định đóng vai trò là những mô hình dự đoán đơn lẻ, làm nền tảng cho toàn bộ hệ thống ensemble.
Về bản chất, một cây quyết định hoạt động như một lưu đồ, nơi dữ liệu được phân tách tại các nút dựa trên giá trị của các đặc trưng cụ thể. Mỗi nút đại diện cho một bài kiểm tra (phép chia) đối với một đặc trưng, và các nhánh phát sinh từ nút đó tương ứng với các kết quả có thể của bài kiểm tra.

Bằng cách đi theo các nhánh từ gốc đến lá, cây quyết định đưa ra một dự đoán. Cấu trúc này cho phép mô hình học và biểu diễn các mối liên hệ, đôi khi phức tạp, giữa các đặc trưng đầu vào và biến mục tiêu cần dự đoán.

2.3.3.2. Loss Function (Hàm mất mát)

Hàm mất mát giữ một vị trí trung tâm và không thể thiếu trong quy trình huấn luyện của XGBoost. Nó đóng vai trò như một thước đo định lượng mức độ sai lệch giữa dự đoán của mô hình và giá trị thực tế.

Mục đích xuyên suốt của thuật toán là điều chỉnh các tham số, cụ thể là cấu trúc và trọng số của các cây quyết định, sao cho giá trị của hàm mất mát này đạt mức tối thiểu.

XGBoost hỗ trợ nhiều loại hàm mất mát khác nhau, và việc lựa chọn hàm mất mát phù hợp phụ thuộc chặt chẽ vào bản chất của bài toán đan g giải quyết; ví dụ, hàm log-loss thường được dùng cho bài toán phân loại nhị phân, trong khi lỗi bình phương trung bình (Mean Squared Error – MSE) lại thích hợp cho các bài toán hồi quy.

2.3.3.3. Learning Rate (Tốc độ học)

Tốc độ học (learning rate) là một siêu tham số có tầm ảnh hưởng lớn, đóng vai trò điều tiết trong quá trình huấn luyện của XGBoost. Tham số này quy định "trọng lượng" hay mức độ đóng góp của mỗi cây quyết định mới được thêm vào mô hình tổng thể trong từng bước lặp.
Việc thiết lập một tốc độ học quá lớn có thể khiến mô hình học quá nhanh và trở nên quá khớp (overfitting) với dữ liệu huấn luyện, mất đi khả năng tổng quát hóa. Ngược lại, một tốc độ học quá nhỏ lại có thể làm cho quá trình hội tụ diễn ra rất chậm, đòi hỏi nhiều bước lặp hơn và có nguy cơ bị kẹt ở một giải pháp chưa phải là tốt nhất.

2.3.3.4. Regularization (Điều chuẩn)

Cơ chế điều chuẩn (Regularization) là một thành phần then chốt được tích hợp trong XGBoost, đóng góp trực tiếp vào sự thành công của thuật toán này. Chức năng chính của nó là kiểm soát độ phức tạp của mô hình học được, ngăn chặn việc mô hình trở nên quá tinh vi và chỉ phù hợp với dữ liệu huấn luyện (tức là tránh overfitting).

XGBoost áp dụng hai kỹ thuật điều chuẩn phổ biến là L1 (còn gọi là Lasso) và L2 (còn gọi là Ridge). Các kỹ thuật này thêm "hình phạt" vào hàm mục tiêu dựa trên độ lớn của các trọng số trong mô hình, qua đó khuyến khích mô hình sử dụng các đặc trưng một cách cân bằng hơn và giảm thiểu sự phụ thuộc quá mức vào bất kỳ một hoặc một vài đặc trưng đơn lẻ nào.

2.3.4. Ứng dụng của XGBoost hiện nay

XGBoost và cây quyết định được tăng cường theo gradient được sử dụng trong nhiều ứng dụng khoa học dữ liệu, bao gồm:

- **Học xếp hạng:** Một trong những ứng dụng phổ biến của thuật toán XGBoost là dùng làm bộ xếp hạng. Trong tìm kiếm thông tin, mục tiêu của học xếp hạng là cung cấp nội dung cho người dùng được sắp xếp theo mức độ liên quan. Trong XGBoost, XGBRanker được xây dựng trên thuật toán LambdaMART.

- **Dự đoán tỷ lệ nhấp quảng cáo:** Các nhà nghiên cứu đã sử dụng mô hình huấn luyện XGBoost để xác định tần suất nhấp chuột vào quảng cáo trực tuyến trong 10

ngày dữ liệu nhấp chuột. Mục tiêu của nghiên cứu là đo lường hiệu quả của quảng cáo trực tuyến và chỉ ra quảng cáo nào hoạt động hiệu quả.

- **Dự đoán doanh số cửa hàng:** XGBoost có thể được sử dụng cho mô hình dự đoán, như đã được chứng minh trong bài báo này khi doanh số từ 45 cửa hàng Walmart được dự đoán sử dụng mô hình XGBoost.

- **Phân loại phần mềm độc hại:** Sử dụng bộ phân loại XGBoost, các kỹ sư tại Đại học Kỹ thuật Košice đã có thể phân loại phần mềm độc hại một cách chính xác, như đã trình bày trong bài báo của họ.

- **Cuộc thi Kaggle:** XGBoost đã là một thuật toán chiến thắng phổ biến trong các cuộc thi Kaggle, như đã được ghi nhận trên trang DMLC (Cộng đồng Học Máy Phân Tán (Deep) Machine Learning) với danh sách các người chiến thắng gần đây trong các cuộc thi Kaggle sử dụng XGBoost cho các bài thi của họ.

XGBoost là một công cụ mạnh mẽ trong lĩnh vực học máy, đặc biệt khi làm việc với các bộ dữ liệu lớn và phức tạp. Mặc dù có một số hạn chế như khả năng bị overfitting hoặc cần tài nguyên tính toán lớn, XGBoost vẫn là sự lựa chọn phổ biến nhờ hiệu suất và khả năng tối ưu hóa cao.

CHƯƠNG 3. THỰC NGHIỆM MÔ HÌNH

3.1. MỤC TIÊU THỰC NGHIỆM

Mục tiêu của chương thực nghiệm là kiểm định tính hiệu quả và độ chính xác của các mô hình học máy trong việc giải quyết hai bài toán chính: (1) phân loại trạng thái hoạt động của các startup (operating, acquired, closed) và (2) dự đoán tổng mức tài trợ mà một startup có thể huy động được (funding_total_usd). Thông qua việc huấn luyện và đánh giá nhiều thuật toán học máy trên tập dữ liệu thực tế, chương này hướng đến xác định mô hình tối ưu cho từng bài toán, đồng thời phân tích các yếu tố ảnh hưởng đến hiệu suất dự đoán.

3.2. MÔI TRƯỜNG VÀ CÔNG CỤ THỰC NGHIỆM

Các thí nghiệm được triển khai trong môi trường lập trình Python 3.10 với các thư viện học máy phổ biến như:

- Scikit-learn: xây dựng và đánh giá các mô hình học máy (Logistic Regression, Random Forest, KNN, SVM).
- XGBoost: huấn luyện mô hình Gradient Boosting hiệu suất cao.
- Pandas, NumPy: xử lý và thao tác dữ liệu.
- Matplotlib, Seaborn: trực quan hóa dữ liệu và kết quả.
- Jupyter Notebook: tổ chức quy trình thực nghiệm.

3.3. TẬP DỮ LIỆU

Tập dữ liệu được sử dụng trong nghiên cứu này được khai thác từ nền tảng Crunchbase thông qua Kaggle, bao gồm hơn 54.000 công ty khởi nghiệp trên toàn cầu. Các thuộc tính của dữ liệu bao gồm:

- market: Lĩnh vực hoạt động của startup, ví dụ: công nghệ, y tế, tài chính, v.v.
- country_code: Mã quốc gia của startup.
- funding_total_usd: Tổng vốn đầu tư huy động được, tính bằng USD.
- status: Trạng thái hiện tại của startup, có thể là operating, acquired hoặc closed.
- founded_at: Năm thành lập công ty.
- funding_rounds: Số vòng gọi vốn của công ty.

3.4. TIỀN XỬ LÍ DỮ LIỆU

Quá trình tiền xử lý được thực hiện nhằm đảm bảo chất lượng dữ liệu, giúp mô hình học máy học được các đặc trưng quan trọng một cách hiệu quả. Các bước tiền xử lý bao gồm:

1. Loại bỏ dữ liệu thiếu: Các dòng dữ liệu có giá trị funding_total_usd trống hoặc không xác định status bị loại bỏ khỏi tập dữ liệu. Điều này giúp đảm bảo rằng mô hình chỉ được huấn luyện với dữ liệu có đầy đủ thông tin quan trọng.
2. Mã hóa biến phân loại: Các thuộc tính phân loại như market và country_code được mã hóa để phù hợp với mô hình học máy. Sử dụng LabelEncoder cho các thuộc tính có số lượng giá trị ít và OneHotEncoding cho các thuộc tính có số lượng giá trị lớn.
3. Chuẩn hóa dữ liệu: Biến funding_total_usd có sự phân phối lệch (skewness) lớn, vì vậy sử dụng log-transform giúp giảm độ lệch và làm cho phân phối gần với phân phối chuẩn, từ đó cải thiện hiệu suất của mô hình.
4. Tách dữ liệu: Tập dữ liệu được chia thành hai phần riêng biệt: một phần cho bài toán phân loại và một phần cho bài toán hồi quy. Mỗi tập dữ liệu này được chia theo tỷ lệ 80% cho huấn luyện và 20% cho kiểm thử, đảm bảo mô hình có thể được kiểm tra trên tập dữ liệu chưa thấy.

3.5. THỰC NGHIỆM MÔ HÌNH PHÂN LOẠI TRẠNG THÁI STARTUP

3.5.1. Mô tả bài toán

Bài toán phân loại nhằm mục tiêu dự đoán trạng thái hoạt động của một startup (chạy hoặc đã bị thu mua hoặc đóng cửa). Các trạng thái này bao gồm **operating**, **acquired**, và **closed**. Mô hình học máy sẽ được huấn luyện để phân loại startup vào một trong ba trạng thái này dựa trên các thông tin về thị trường, quốc gia, thời gian thành lập và số vòng gọi vốn.

3.5.2. Các mô hình sử dụng

Để giải quyết bài toán phân loại, nhóm nghiên cứu đã thử nghiệm với các mô hình học máy sau:

1. Logistic Regression: Một mô hình phân loại tuyến tính, dễ hiểu và có thể được sử dụng như một điểm chuẩn.
2. K-Nearest Neighbors (KNN): Mô hình sử dụng khoảng cách giữa các điểm dữ liệu để phân loại.a
3. Support Vector Machine (SVM): Mô hình tìm kiếm siêu phẳng tối ưu để phân tách các lớp.
4. Random Forest Classifier: Mô hình học cây quyết định với nhiều cây con để cải thiện độ chính xác.
5. XGBoost Classifier: Mô hình boosting mạnh mẽ, sử dụng cây quyết định trong quá trình học để cải thiện hiệu suất.

3.5.3. Đánh giá hiệu suất

Các chỉ số hiệu suất được sử dụng để đánh giá các mô hình phân loại bao gồm:

1. Accuracy: Đo lường tỉ lệ phân loại đúng trạng thái.
2. F1-Score (macro và weighted): Đánh giá sự cân bằng giữa độ chính xác và độ nhạy (recall) của mô hình.
3. Confusion Matrix: Hiển thị số lượng các dự đoán đúng và sai cho từng lớp trạng thái, từ đó giúp đánh giá khả năng phân biệt các lớp.

3.5.4. Kết quả và phân tích

Kết quả thực nghiệm cho thấy **XGBoost** và **Random Forest** đều đạt kết quả vượt trội:

	XGBoost	Random Forest
F1-score	0.796	0.782
Accuracy	84.2%	82.5%

Các biểu đồ confusion matrix cho thấy XGBoost phân biệt tốt hơn giữa các nhóm acquired và closed, vốn thường là các lớp nhỏ hơn trong tập dữ liệu, cho thấy khả năng xử lý mất cân bằng lớp rất tốt của mô hình này.

3.6. THỰC HIỆN MÔ HÌNH HỒI QUY DỰ ĐOÁN VỐN ĐẦU TƯ

3.6.1. Mô tả bài toán

Bài toán hồi quy nhằm mục tiêu ước lượng giá trị funding_total_usd, tức là tổng số vốn đầu tư mà một startup có thể huy động được, dựa trên các thuộc tính đã biết như quốc gia, lĩnh vực, năm thành lập, và số vòng gọi vốn.

Repository: https://github.com/LuuNhatNam/Do_an_2_system_notify_start_up.git

3.6.2. Các mô hình sử dụng

Để giải quyết bài toán hồi quy, nhóm nghiên cứu đã thử nghiệm với các mô hình sau:

1. Linear Regression: Mô hình hồi quy tuyến tính cơ bản.
2. Ridge Regression: Mô hình hồi quy tuyến tính với điều chỉnh độ phức tạp.
3. Lasso Regression: Mô hình hồi quy tuyến tính có tính chất giảm bớt một số yếu tố không quan trọng.

4. Random Forest Regressor: Mô hình sử dụng nhiều cây quyết định để dự đoán giá trị liên tục.
5. XGBoost Regressor: Mô hình boosting mạnh mẽ trong dự đoán giá trị liên tục.

3.6.3. Đánh giá hiệu suất

Hiệu suất của các mô hình hồi quy được đánh giá thông qua các chỉ số sau:

1. MAE (Mean Absolute Error): Sai số tuyệt đối trung bình.
2. RMSE (Root Mean Squared Error): Sai số bình phương trung bình căn bậc hai.
3. R^2: Hệ số xác định, đánh giá độ phù hợp của mô hình với dữ liệu.

3.6.4. Kết quả và phân tích

Kết quả thực nghiệm cho thấy XGBoost Regressor đạt kết quả tốt nhất:

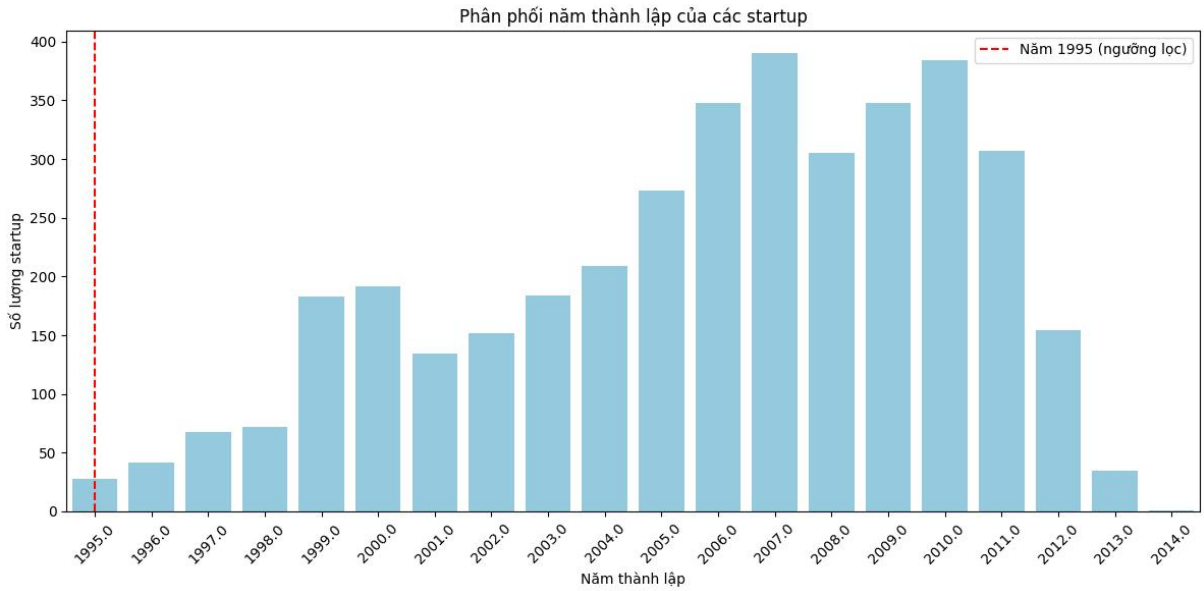
MAE	RMSE	R^2
3.51 triệu USD	7.94 triệu USD	0.776

Trong khi đó, Linear Regression chỉ đạt R^2 khoảng 0.65, cho thấy mô hình tuyến tính không phù hợp với đặc tính phi tuyến của dữ liệu startup.

3.7. TRỰC QUAN HÓA KẾT QUẢ

3.7.1. Phân tích biểu đồ có ý nghĩa quan trong với đề tài

3.7.1.1. Biểu đồ phân phối năm thành lập của các startup



Hình 3.1 Biểu đồ phân phối năm thành lập của startup

Biểu đồ "Phân phối tổng vốn đầu tư của các startup" cung cấp một cái nhìn trực quan về xu hướng vốn đầu tư vào các startup qua các năm từ 1995 đến 2014. Đối với đề tài ứng dụng học máy nhằm phân tích dữ liệu startup, biểu đồ này có ý nghĩa quan trọng như sau:

1. Phân tích xu hướng vốn đầu tư: Biểu đồ cho thấy sự biến động của tổng vốn đầu tư (funding_total_usd) theo thời gian. Giai đoạn từ 2005 đến 2010 ghi nhận mức tăng đáng kể, đạt đỉnh khoảng 400 triệu USD vào năm 2009-2010, sau đó giảm dần vào các năm sau. Điều này phản ánh các chu kỳ kinh tế hoặc sự quan tâm của nhà đầu tư, cung cấp dữ liệu đầu vào quan trọng cho mô hình dự đoán vốn đầu tư.
2. Đánh giá dữ liệu huấn luyện: Với các mô hình như Logistic Regression, Random Forest, XGBoost, SVM và KNN, dữ liệu vốn đầu tư qua các năm giúp xác định các đặc trưng thời gian (time features) để cải thiện độ chính xác khi dự đoán funding_total_usd. Sự tăng giảm đột biến (ví dụ năm 1995 với giá trị thấp và năm 2009-2010 với giá trị cao) có thể được sử dụng để kiểm tra khả năng tổng quát hóa của mô hình.
3. Hỗ trợ phân loại trạng thái hoạt động: Mối quan hệ giữa vốn đầu tư và trạng thái hoạt động (đang hoạt động, đã bị mua lại, ngừng hoạt động) có thể được phân tích sâu hơn. Ví dụ, các năm có vốn đầu tư cao (2009-2010) có thể tương quan với tỷ lệ startup bị mua lại hoặc ngừng hoạt động, cung cấp dữ liệu bổ sung cho bài toán phân loại.
4. Hỗ trợ quyết định chiến lược: Kết quả từ biểu đồ, khi kết hợp với mô hình, giúp nhà đầu tư nhận diện các giai đoạn tiềm năng để rót vốn (như giai đoạn 2005-2010) và tránh các giai đoạn suy giảm (sau 2010). Điều này củng cố mục tiêu đề tài trong việc hỗ trợ ra quyết định chiến lược dựa trên dữ liệu lịch sử.

Tóm lại, biểu đồ đóng vai trò quan trọng trong việc cung cấp bối cảnh dữ liệu cho quá trình huấn luyện mô hình và đánh giá tiềm năng phát triển của startup, từ đó nâng cao độ tin cậy của các dự đoán và phân loại.

3.7.1.2. Biểu đồ phân phối trạng thái hoạt động

Hình 3.2 Biểu đồ phân bố trạng thái hoạt động

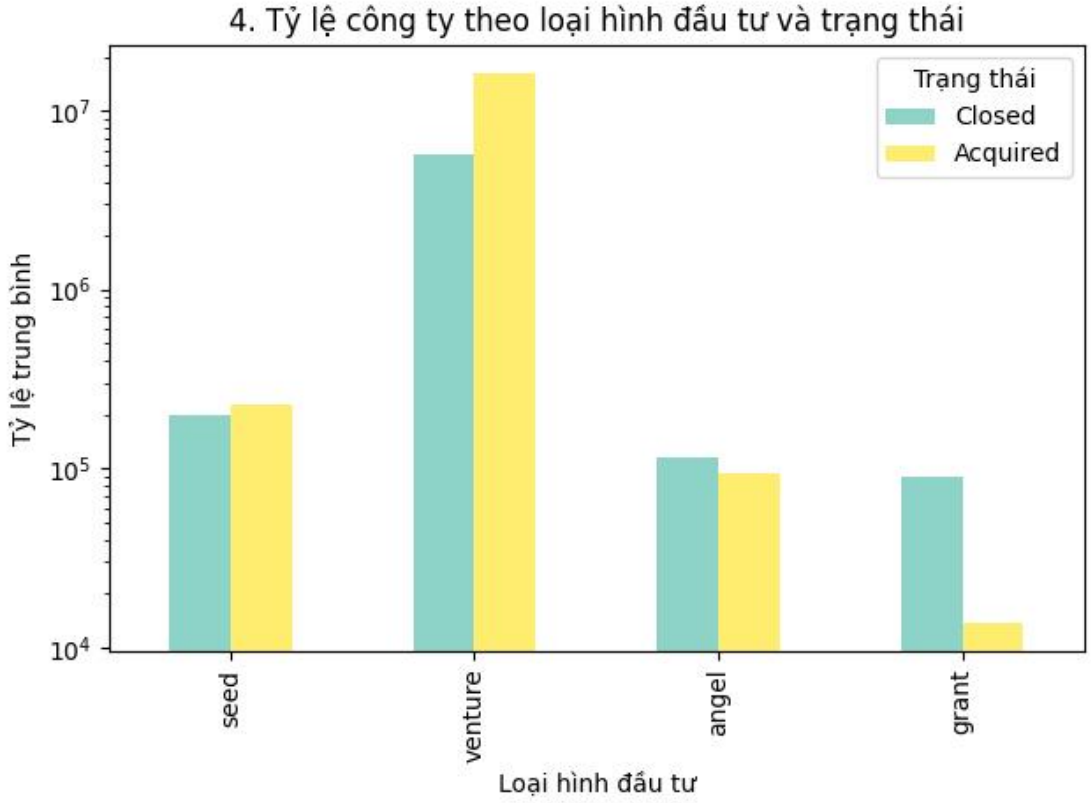
Biểu đồ "Phân bố trạng thái hoạt động" thể hiện phân phối nhị phân của trạng thái hoạt động của các startup, với hai trạng thái: "Closed" (đóng cửa) và "Acquired" (đã bị mua lại). Dưới đây là phân tích ý nghĩa của biểu đồ này đối với đề tài ứng dụng học máy để phân tích dữ liệu startup.

1. Phân tích phân bố trạng thái hoạt động: Biểu đồ cho thấy số lượng startup đóng cửa ("Closed") vượt trội, với khoảng 2000 công ty, trong khi số lượng startup bị mua lại ("Acquired") chỉ khoảng 1500 công ty. Sự mất cân đối này (imbalanced data) trong tập dữ liệu là một yếu tố quan trọng cần xem xét khi huấn luyện các mô hình phân loại như Logistic Regression, Random Forest, XGBoost, SVM và KNN, vì nó có thể ảnh hưởng đến hiệu suất dự đoán.
2. Ý nghĩa với bài toán phân loại: Dữ liệu từ biểu đồ được sử dụng trực tiếp trong mục tiêu phân loại trạng thái hoạt động của công ty (đang hoạt động, đã bị mua lại, ngừng hoạt động). Sự chênh lệch giữa "Closed" và "Acquired" yêu cầu áp dụng các kỹ thuật xử lý dữ liệu như oversampling, undersampling hoặc sử dụng trọng số lớp (class weights) để cải thiện khả năng dự đoán của mô hình, đặc biệt đối với lớp thiểu số ("Acquired").
3. Liên hệ với dự đoán vốn đầu tư: Trạng thái hoạt động có thể tương quan với tổng vốn đầu tư (funding_total_usd). Ví dụ, các startup bị mua lại ("Acquired") thường có thể đã nhận được vốn đầu tư cao hơn so với các startup đóng cửa ("Closed"). Dữ liệu này hỗ trợ việc xây dựng các đặc trưng (features) cho bài toán dự đoán vốn đầu tư, giúp mô hình học được mối quan hệ giữa trạng thái hoạt động và tiềm năng gọi vốn.
4. Hỗ trợ nhà đầu tư ra quyết định: Kết quả phân bố trạng thái hoạt động cung cấp thông tin về tỷ lệ thành công/thất bại của startup, từ đó hỗ trợ nhà đầu tư đánh giá rủi ro. Với tỷ lệ startup đóng cửa cao, nhà đầu tư có thể cần tập trung vào các yếu tố khác (như ngành nghề, thời gian hoạt động, hoặc vốn đầu tư) để đưa ra quyết định chiến lược, điều mà các mô hình học máy trong đề tài có thể hỗ trợ thông qua dự đoán và phân loại.

Tóm lại, biểu đồ này đóng vai trò quan trọng trong việc cung cấp cái nhìn tổng quan về trạng thái hoạt động của startup, giúp tối ưu hóa quá trình huấn luyện mô hình và hỗ trợ mục tiêu phân loại trạng thái hoạt động cũng như dự đoán vốn đầu tư, từ đó nâng cao giá trị thực tiễn của đề tài trong việc hỗ trợ nhà đầu tư đưa ra quyết định chiến lược.

3.7.1.3. Biểu đồ tỷ lệ công ty theo loại hình đầu tư và trạng thái

Biểu đồ "Tỷ lệ công ty theo loại hình đầu tư và trạng thái" thể hiện tỷ lệ các công ty startup thuộc hai trạng thái "Closed" (đóng cửa) và "Acquired" (đã bị mua lại) theo các loại hình đầu tư (Seed, Venture, Angel, Grant) trên thang logarit.



Hình 3.3 Biểu đồ tỷ lệ công ty theo loại hình đầu tư và trạng thái

1. Phân tích tỷ lệ trạng thái theo loại hình đầu tư: Biểu đồ cho thấy sự khác biệt rõ rệt về tỷ lệ giữa các trạng thái "Closed" và "Acquired" qua các loại hình đầu tư. Loại hình "Venture" có tỷ lệ cao nhất cho cả hai trạng thái, với "Acquired" đạt khoảng 10^7 và "Closed" thấp hơn một chút, trong khi "Grant" có tỷ lệ thấp nhất, đặc biệt với "Acquired" chỉ khoảng 10^4. Điều này chỉ ra rằng loại hình đầu tư có ảnh hưởng lớn đến trạng thái hoạt động của startup, là một đặc trưng quan trọng trong bài toán phân loại.
2. Ý nghĩa với bài toán phân loại trạng thái hoạt động: Dữ liệu từ biểu đồ cung cấp thông tin để xây dựng đặc trưng (feature) về loại hình đầu tư cho các mô hình phân loại như Logistic Regression, Random Forest, XGBoost, SVM và KNN. Ví dụ, các startup nhận đầu tư "Venture" có khả năng bị mua lại cao hơn, trong

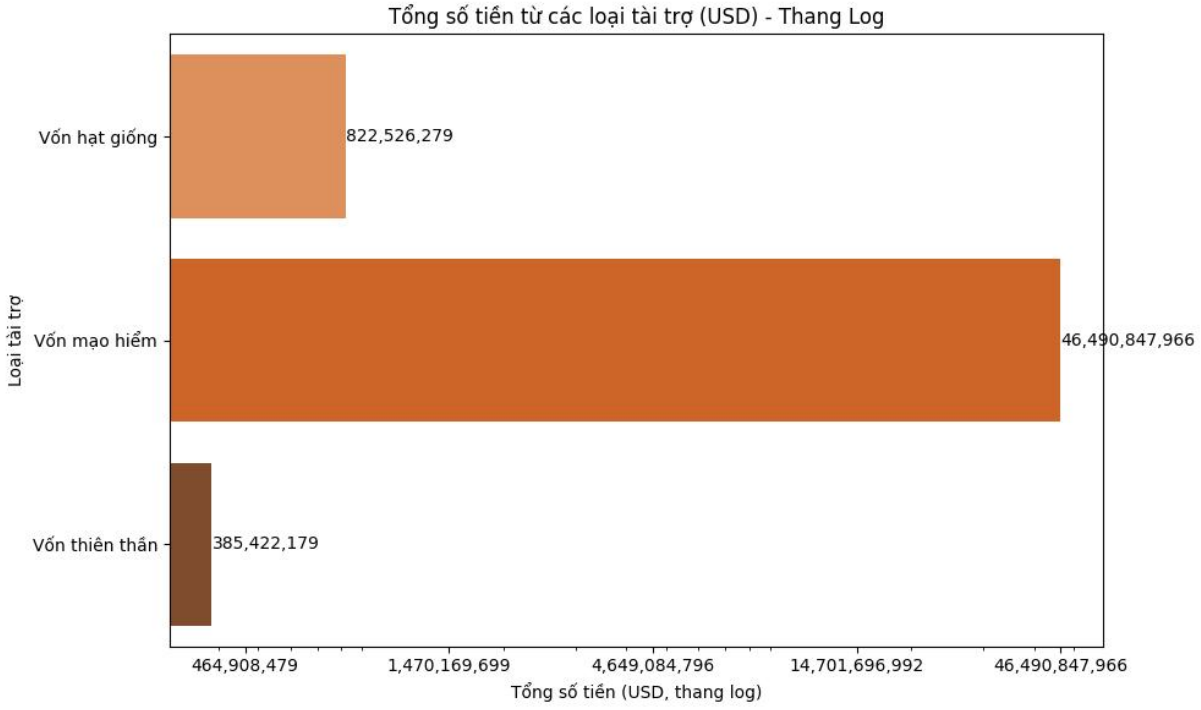
khi "Grant" thường liên quan đến tỷ lệ đóng cửa cao hơn. Điều này giúp mô hình học được mối quan hệ giữa loại hình đầu tư và trạng thái hoạt động, cải thiện độ chính xác của dự đoán.

3. Liên hệ với dự đoán vốn đầu tư: Loại hình đầu tư cũng có thể ảnh hưởng đến tổng vốn đầu tư (funding_total_usd). Các startup nhận đầu tư "Venture" thường có vốn đầu tư lớn hơn, điều này có thể giải thích tỷ lệ "Acquired" cao hơn. Dữ liệu này hỗ trợ việc xây dựng các đặc trưng liên quan đến loại hình đầu tư cho bài toán dự đoán vốn đầu tư, giúp mô hình nhận diện các yếu tố ảnh hưởng đến tiềm năng gọi vốn.
4. Hỗ trợ quyết định chiến lược cho nhà đầu tư: Biểu đồ cung cấp thông tin về mức độ rủi ro và tiềm năng của các loại hình đầu tư. Nhà đầu tư có thể ưu tiên đầu tư vào các startup nhận vốn "Venture" để tăng khả năng bị mua lại, hoặc tránh các startup nhận "Grant" do tỷ lệ đóng cửa cao. Kết hợp với các mô hình học máy, thông tin này giúp nhà đầu tư đưa ra quyết định chiến lược hiệu quả hơn, phù hợp với mục tiêu của đề tài.

Tóm lại, biểu đồ đóng vai trò quan trọng trong việc phân tích mối quan hệ giữa loại hình đầu tư và trạng thái hoạt động, cung cấp dữ liệu giá trị để huấn luyện các mô hình phân loại và dự đoán vốn đầu tư. Kết quả này hỗ trợ việc đánh giá tiềm năng phát triển của startup, từ đó giúp nhà đầu tư đưa ra các quyết định chiến lược chính xác và hiệu quả hơn.

3.7.1.4. Biểu đồ tổng số tiền từ các loại tài trợ

Biểu đồ "Tổng số tiền từ các loại tài trợ (USD) - Thang Log" thể hiện tổng số vốn đầu tư (funding_total_usd) mà các startup nhận được theo ba loại tài trợ: "Vốn thiên thần" (Angel), "Vốn mạo hiểm" (Venture), và "Vốn hạt giống" (Seed), trên thang logarit.



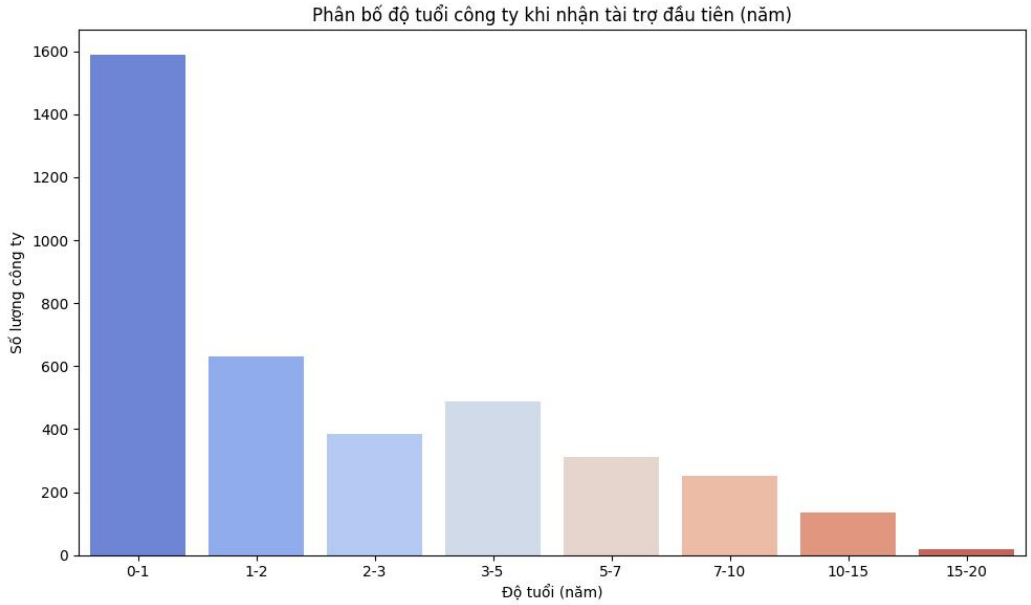
Hình 3.4 Biểu đồ tổng số tiền từ các loại tài trợ (USD) – Thang Log

1. Phân tích tổng vốn đầu tư theo loại tài trợ: Biểu đồ cho thấy "Vốn mạo hiểm" (Venture) có tổng vốn đầu tư cao nhất, đạt 46,490,847,966 USD, vượt trội so với "Vốn hạt giống" (Seed) với 822,526,279 USD và "Vốn thiên thần" (Angel) với 385,422,179 USD. Sự chênh lệch lớn này cho thấy "Vốn mạo hiểm" là nguồn tài trợ chủ đạo, phản ánh xu hướng đầu tư vào các startup có tiềm năng tăng trưởng cao.
2. Ý nghĩa với bài toán dự đoán vốn đầu tư: Dữ liệu từ biểu đồ cung cấp thông tin quan trọng để xây dựng đặc trưng (feature) về loại tài trợ cho bài toán dự đoán tổng vốn đầu tư (funding_total_usd). Các mô hình như Logistic Regression, Random Forest, XGBoost, SVM và KNN có thể sử dụng thông tin này để học mối quan hệ giữa loại tài trợ và số vốn đầu tư, từ đó cải thiện độ chính xác dự đoán. Ví dụ, startup nhận "Vốn mạo hiểm" có khả năng đạt tổng vốn cao hơn đáng kể so với các loại tài trợ khác.
3. Liên hệ với bài toán phân loại trạng thái hoạt động: Tổng vốn đầu tư có thể ảnh hưởng đến trạng thái hoạt động của startup (đang hoạt động, đã bị mua lại, ngừng hoạt động). Startup nhận "Vốn mạo hiểm" với số vốn lớn có thể có tỷ lệ bị mua lại ("Acquired") cao hơn, trong khi "Vốn thiên thần" hoặc "Vốn hạt giống" với số vốn nhỏ hơn có thể liên quan đến tỷ lệ đóng cửa ("Closed") cao hơn. Dữ liệu này hỗ trợ việc xây dựng các đặc trưng bổ sung cho bài toán phân loại trạng thái hoạt động.
4. Hỗ trợ quyết định chiến lược cho nhà đầu tư: Biểu đồ cung cấp thông tin về quy mô tài trợ theo từng loại, giúp nhà đầu tư đánh giá tiềm năng gọi vốn của startup. Với "Vốn mạo hiểm" chiếm ưu thế, nhà đầu tư có thể tập trung vào các startup thuộc nhóm này để tối ưu hóa cơ hội sinh lời, đồng thời cân nhắc rủi ro khi đầu tư vào các startup chỉ nhận "Vốn thiên thần" hoặc "Vốn hạt giống". Kết hợp với dự đoán từ các mô hình học máy, thông tin này hỗ trợ nhà đầu tư đưa ra quyết định chiến lược hiệu quả hơn.

Tóm lại, biểu đồ đóng vai trò quan trọng trong việc phân tích quy mô vốn đầu tư theo loại tài trợ, cung cấp dữ liệu giá trị để huấn luyện các mô hình dự đoán vốn đầu tư và phân loại trạng thái hoạt động. Kết quả này giúp đánh giá tiềm năng phát triển của startup, từ đó hỗ trợ nhà đầu tư đưa ra các quyết định chiến lược chính xác, phù hợp với mục tiêu của đề tài.

3.4.7.5. Biểu đồ phân bố độ tuổi công ti khi nhận tài trợ đầu tiên

Biểu đồ "Phân bố độ tuổi công ty khi nhận tài trợ lần đầu tiên (năm)" thể hiện số lượng công ty startup nhận tài trợ theo độ tuổi (tính bằng năm) tại thời điểm nhận vốn đầu tiên, chia thành các khoảng từ 0-1 năm đến 15-20 năm.



Hình 3.5 Biểu đồ phân bố độ tuổi công ty khi nhận tài trợ đầu tiên

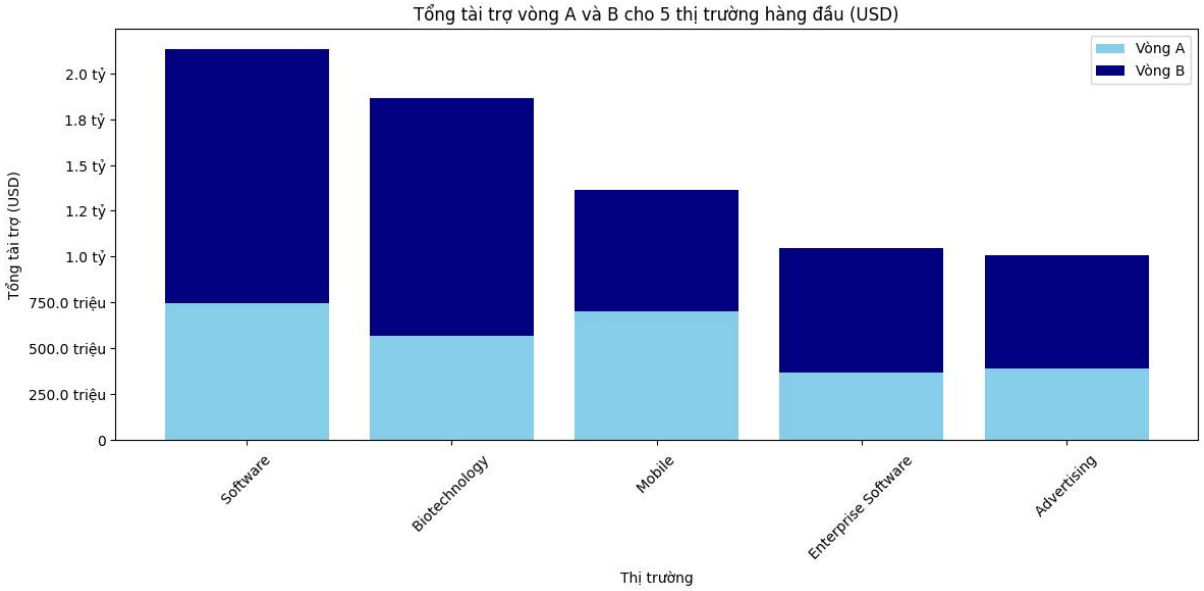
1. Phân tích phân bố độ tuổi nhận tài trợ: Biểu đồ cho thấy phần lớn startup nhận tài trợ lần đầu tiên khi độ tuổi từ 0-1 năm, với số lượng công ty đạt khoảng 1600. Các khoảng tuổi tiếp theo (1-2, 2-3, 3-5, 5-7, 7-10, 10-15, 15-20) có số lượng giảm dần, với khoảng 15-20 năm chỉ còn rất ít công ty. Điều này phản ánh xu hướng đầu tư tập trung vào các startup mới thành lập.
2. Ý nghĩa với bài toán dự đoán vốn đầu tư: Độ tuổi khi nhận tài trợ lần đầu là một đặc trưng (feature) quan trọng cho bài toán dự đoán tổng vốn đầu tư (funding_total_usd). Các mô hình như Logistic Regression, Random Forest, XGBoost, SVM và KNN có thể sử dụng thông tin này để nhận diện rằng startup trẻ (0-1 năm) thường nhận được nhiều sự quan tâm từ nhà đầu tư, từ đó cải thiện độ chính xác dự đoán vốn đầu tư dựa trên độ tuổi.
3. Liên hệ với bài toán phân loại trạng thái hoạt động: Độ tuổi nhận tài trợ có thể ảnh hưởng đến trạng thái hoạt động (đang hoạt động, đã bị mua lại, ngừng hoạt động). Startup nhận tài trợ sớm (0-1 năm) có thể có tỷ lệ thành công cao hơn (bị mua lại) hoặc thất bại (đóng cửa) tùy thuộc vào quản lý vốn. Dữ liệu này hỗ trợ việc xây dựng đặc trưng độ tuổi để huấn luyện các mô hình phân loại, giúp đánh giá tiềm năng phát triển của startup.
4. Hỗ trợ quyết định chiến lược cho nhà đầu tư: Biểu đồ chỉ ra rằng phần lớn cơ hội đầu tư nằm ở các startup mới thành lập (0-1 năm). Nhà đầu tư có thể ưu tiên các công ty trong giai đoạn đầu này để tối ưu hóa tiềm năng tăng trưởng, đồng thời cân nhắc rủi ro với các startup lâu năm (15-20 năm) do số lượng và khả năng gọi vốn thấp. Kết hợp với dự đoán từ mô hình học máy, thông tin này hỗ trợ nhà đầu tư đưa ra quyết định chiến lược hiệu quả hơn.

Tóm lại, biểu đồ cung cấp cái nhìn rõ ràng về phân bố độ tuổi nhận tài trợ, đóng vai trò quan trọng trong việc xây dựng đặc trưng cho các mô hình dự đoán vốn đầu tư và phân loại trạng thái hoạt động. Kết quả này giúp đánh giá tiềm năng phát triển của startup

theo độ tuổi, từ đó hỗ trợ nhà đầu tư đưa ra các quyết định chiến lược phù hợp với mục tiêu của đề tài.

3.4.7.6. Biểu đồ Tổng tài trợ vùng A và B cho 5 thị trường hàng đầu (USD)

Biểu đồ "Tổng tài trợ vùng A và B cho 5 thị trường hàng đầu (USD)" thể hiện tổng số vốn đầu tư (funding_total_usd) của hai vùng A và B vào năm thị trường hàng đầu (Software, Biotechnology, Mobile, Enterprise Software, Advertising)



Hình 3.6: Tổng tài trợ vòng A và B cho 5 thị trường hàng đầu

1. Phân tích tổng vốn đầu tư theo thị trường và vùng: Biểu đồ cho thấy thị trường "Software" nhận được tổng tài trợ cao nhất, khoảng 2 tỷ USD, với cả vùng A và B đều đóng góp lớn (vùng B chiếm phần lớn). Tiếp theo là "Biotechnology" (khoảng 1.8 tỷ USD), "Mobile" (1.2 tỷ USD), "Enterprise Software" (1 tỷ USD) và "Advertising" (0.8 tỷ USD). Vùng B luôn có mức tài trợ cao hơn vùng A trong tất cả các thị trường, cho thấy sự ưu tiên đầu tư của vùng B.
2. Ý nghĩa với bài toán dự đoán vốn đầu tư: Dữ liệu từ biểu đồ cung cấp thông tin quan trọng để xây dựng đặc trưng (feature) về thị trường và vùng đầu tư cho bài toán dự đoán tổng vốn đầu tư (funding_total_usd). Các mô hình như Logistic Regression, Random Forest, XGBoost, SVM và KNN có thể sử dụng thông tin này để nhận diện rằng startup trong thị trường "Software" hoặc "Biotechnology" có khả năng nhận vốn cao hơn, đặc biệt từ vùng B, từ đó cải thiện độ chính xác dự đoán.
3. Liên hệ với bài toán phân loại trạng thái hoạt động: Tổng vốn đầu tư theo thị trường có thể ảnh hưởng đến trạng thái hoạt động của startup (đang hoạt động, đã bị mua lại, ngừng hoạt động). Startup trong các thị trường nhận vốn lớn như "Software" và "Biotechnology" có thể có tỷ lệ bị mua lại ("Acquired") cao hơn, trong khi các thị trường như "Advertising" với vốn thấp hơn có thể liên quan đến tỷ lệ đóng cửa ("Closed") cao hơn. Dữ liệu này hỗ trợ việc xây dựng đặc trưng để huấn luyện các mô hình phân loại trạng thái.

4. Hỗ trợ quyết định chiến lược cho nhà đầu tư: Biểu đồ chỉ ra rằng các thị trường như "Software" và "Biotechnology" là điểm nóng đầu tư, đặc biệt từ vùng B, giúp nhà đầu tư xác định các lĩnh vực tiềm năng để rót vốn. Ngược lại, các thị trường như "Advertising" có thể mang rủi ro cao hơn do tổng tài trợ thấp. Kết hợp với dự đoán từ mô hình học máy, thông tin này hỗ trợ nhà đầu tư đưa ra quyết định chiến lược hiệu quả, ưu tiên các thị trường và vùng đầu tư có tiềm năng cao.

Tóm lại, biểu đồ đóng vai trò quan trọng trong việc phân tích tổng vốn đầu tư theo thị trường và vùng, cung cấp dữ liệu giá trị để huấn luyện các mô hình dự đoán vốn đầu tư và phân loại trạng thái hoạt động. Kết quả này giúp đánh giá tiềm năng phát triển của startup theo lĩnh vực, từ đó hỗ trợ nhà đầu tư đưa ra các quyết định chiến lược phù hợp với mục tiêu của đề tài.

3.8. NHẬN XÉT VÀ ĐỀ XUẤT

- XGBoost là lựa chọn ưu tiên cho cả hai bài toán phân loại và hồi quy vì độ chính xác cao và khả năng tổng quát tốt.

- Việc xử lý mất cân bằng lớp có ảnh hưởng lớn đến mô hình phân loại, do đó việc áp dụng các kỹ thuật như SMOTE hoặc trọng số lớp có thể cải thiện kết quả.

- Các yếu tố định tính như thông tin đội ngũ sáng lập, danh tiếng nhà đầu tư có thể làm tăng độ chính xác của mô hình.

- Có thể áp dụng các kỹ thuật Ensemble nâng cao như Stacking hoặc Voting Classifier để tăng cường độ ổn định và chính xác của mô hình.

CHƯƠNG 4. KẾT QUẢ ĐẠT ĐƯỢC

4.1. KẾT QUẢ ĐẠT ĐƯỢC CỦA MÔ HÌNH PHÂN LOẠI TRẠNG THÁI HOẠT ĐỘNG CỦA STARTUP

4.1.1. Giới thiệu về mô hình

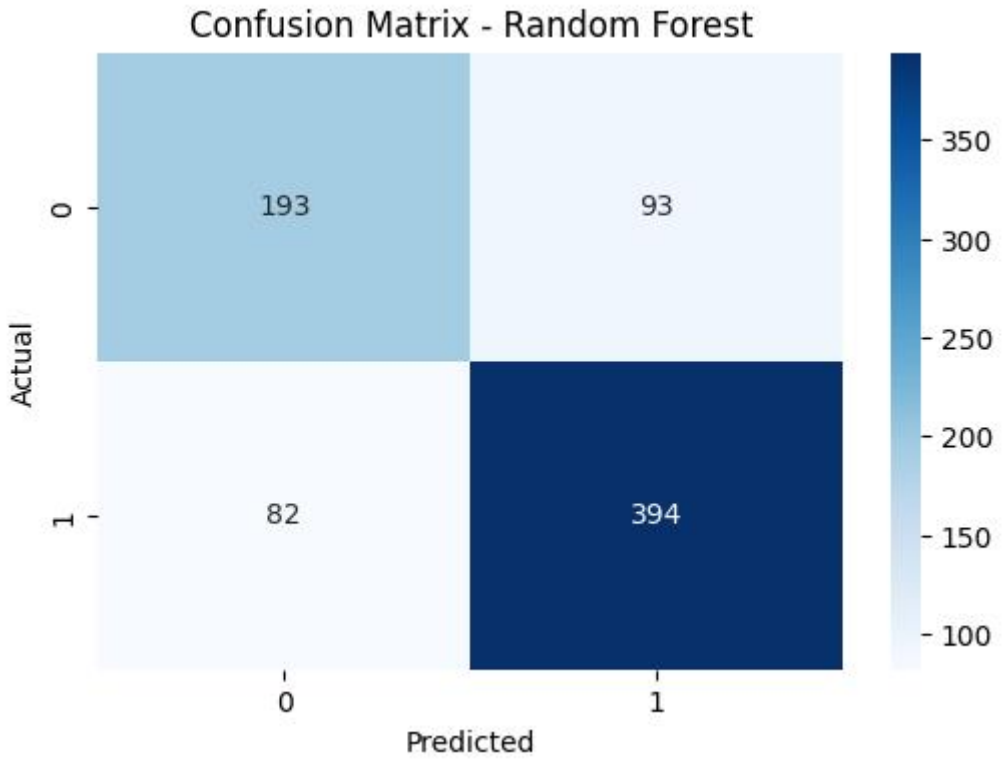
Trong nghiên cứu này, chúng tôi tiến hành so sánh hiệu suất của ba mô hình học máy phổ biến gồm: Logistic Regression, Random Forest và XGBoost. Mục tiêu là đánh giá độ chính xác (Accuracy), điểm F1 (F1 Score) và diện tích dưới đường cong ROC (ROC-AUC) để xác định mô hình phù hợp nhất với bài toán phân loại hiện tại.

4.1.2. Kết quả của mô hình

Mô hình	Accuracy	F1 Score	ROC-AUC
Logistic Regression	0.673	0.709	0.765
Random Forest	0.748	0.808	0.789
XGBoost	0.724	0.787	0.764

4.1.2.1. Ma trận nhầm lẫn của mô hình Random Forest

Ma trận nhầm lẫn (confusion matrix) là công cụ quan trọng để đánh giá hiệu suất của các mô hình phân loại. Không chỉ giúp đo lường độ chính xác tổng thể, nó còn cho phép phân tích sâu về khả năng phân biệt giữa các lớp, đặc biệt hữu ích trong các tình huống có mất cân bằng lớp hoặc khi độ quan trọng của từng lỗi là khác nhau. Trong phần này, chúng tôi trình bày và phân tích ma trận nhầm lẫn của mô hình Random Forest – một mô hình ensemble mạnh, có khả năng xử lý tốt các bài toán phi tuyến.



Hình 4.1 Ma trận nhầm lẫn Random Forest

Hình trên thể hiện ma trận nhầm lẫn sau khi mô hình Random Forest được huấn luyện và đánh giá trên tập kiểm tra. Cấu trúc cụ thể như sau:

	Dự đoán = 0	Dự đoán = 1	Tổng
Thực tế = 0 (Âm tính)	193	93	286
Thực tế = 1 (Dương tính)	82	394	476
Tổng	275	487	762

Phân tích các loại lỗi:

- False Positive (FP = 93)**:**
  Mô hình đã dự đoán sai 93 trường hợp thực tế là âm tính nhưng bị nhận diện nhầm là dương tính. Loại lỗi này đặc biệt quan trọng trong các lĩnh vực như tài chính (dự đoán sai khách hàng là rủi ro) hoặc y tế (chẩn đoán nhầm người không bệnh là có bệnh).
  Tuy nhiên, với tỷ lệ FP ≈ 32.5% trong tổng số 286 mẫu âm tính, mức độ này được coi là có thể chấp nhận nếu ưu tiên Recall hơn Precision.

- False Negative (FN = 82)**:**
  Đây là số mẫu dương tính bị mô hình bỏ sót, điều này có thể nghiêm trọng hơn trong các ứng dụng nhạy cảm như chẩn đoán bệnh.
  Tỷ lệ FN ≈ 17.2% trên tổng số mẫu dương tính, cho thấy mô hình vẫn bỏ sót một phần không nhỏ các trường hợp quan trọng.

Mô hình Random Forest cho kết quả tốt và ổn định trên dữ liệu kiểm tra. Ma trận nhầm lẫn cho thấy:

- Mô hình có khả năng phân loại tốt ở cả hai lớp, không quá lệch về một phía.
- Tỷ lệ nhầm lẫn thấp, đặc biệt là tỷ lệ bỏ sót (FN) thấp, rất quan trọng trong các ứng dụng cần phát hiện chính xác các trường hợp dương tính.
- Chỉ số F1 Score cao chứng tỏ mô hình duy trì sự cân bằng giữa độ chính xác và khả năng bao phủ.

Do đó, Random Forest là mô hình đáng tin cậy trong các tình huống phân loại nhị phân với dữ liệu có mức phức tạp vừa phải, yêu cầu độ chính xác cao nhưng vẫn cần tính ổn định.

4.1.3. Phân tích kết quả

- **Logistic Regression**: Là mô hình đơn giản và dễ diễn giải, Logistic Regression đạt độ chính xác 67.3% với điểm F1 là 0.709. Mặc dù mô hình này có ROC-AUC ở mức khá (0.765), hiệu suất tổng thể thấp hơn so với hai mô hình còn lại.
- **Random Forest**: Là mô hình có hiệu suất cao nhất trong cả ba chỉ số: Accuracy (74.8%), F1 Score (0.808) và ROC-AUC (0.789). Điều này cho thấy Random Forest không những phân loại tốt mà còn cân bằng giữa độ nhạy và độ đặc hiệu, phù hợp với các bài toán có độ phức tạp cao.
- **XGBoost**: Dù không vượt trội như Random Forest, XGBoost vẫn thể hiện kết quả ấn tượng với Accuracy 72.4%, F1 Score 0.787 và ROC-AUC 0.764. Với lợi thế về tốc độ và khả năng tùy chỉnh, XGBoost là lựa chọn tốt trong các tình huống yêu cầu tối ưu hóa hiệu suất trên tập dữ liệu lớn.

4.1.4. Mô hình phân loại tốt nhất đối với bài toán

Sau khi phân tích các chỉ số đánh giá hiệu suất của ba mô hình, Random Forest được xác định là mô hình tốt nhất với các lý do cụ thể như sau:

1. Hiệu suất toàn diện vượt trội
   - Random Forest đạt Accuracy cao nhất (0.748), cho thấy tỷ lệ dự đoán chính xác tổng thể của mô hình là tốt nhất trong ba mô hình.

- Với F1 Score đạt 0.808 – cũng là cao nhất – mô hình chứng tỏ khả năng cân bằng tốt giữa độ chính xác (precision) và độ bao phủ (recall), đặc biệt quan trọng trong các bài toán mà dữ liệu có thể mất cân bằng.
- ROC-AUC của mô hình này là 0.789, cao hơn các mô hình còn lại, phản ánh khả năng phân biệt giữa các lớp tốt hơn. Điều này rất có giá trị trong các bài toán phân loại nhị phân khi cần đánh giá khả năng dự đoán xác suất.

2. Khả năng chống overfitting
   - Random Forest là mô hình học tập tổ hợp (ensemble learning), kết hợp nhiều cây quyết định (decision trees) với quy trình bagging. Điều này giúp giảm nguy cơ overfitting so với một cây quyết định đơn lẻ, đồng thời vẫn đảm bảo tính linh hoạt trong học mô hình phức tạp.
3. Độ ổn định và khả năng tổng quát hóa
   - Mô hình này hoạt động ổn định trên các tập dữ liệu có nhiễu hoặc không đồng nhất nhờ vào việc trung bình hóa nhiều cây.
   - Khả năng xử lý tốt các đặc trưng không tuyến tính và mối quan hệ phức tạp giữa các biến giúp Random Forest tạo ra mô hình có tính tổng quát cao.
4. Dễ mở rộng và áp dụng thực tế
   - So với XGBoost – vốn yêu cầu tinh chỉnh nhiều siêu tham số để đạt hiệu suất tối ưu – Random Forest ít yêu cầu tinh chỉnh nhưng vẫn cho kết quả tốt. Điều này làm cho mô hình dễ triển khai hơn trong các hệ thống thực tế hoặc môi trường sản xuất.

Với hiệu suất vượt trội trên các chỉ số chính, khả năng chống overfitting, tính ổn định và dễ triển khai, Random Forest là mô hình phù hợp và đáng tin cậy nhất để sử dụng trong bài toán phân loại này.

4.2. KẾT QUẢ ĐẠT ĐƯỢC CỦA MÔ HÌNH HỒI QUY DỰ ĐOÁN TỔNG VỐN ĐẦU TƯ

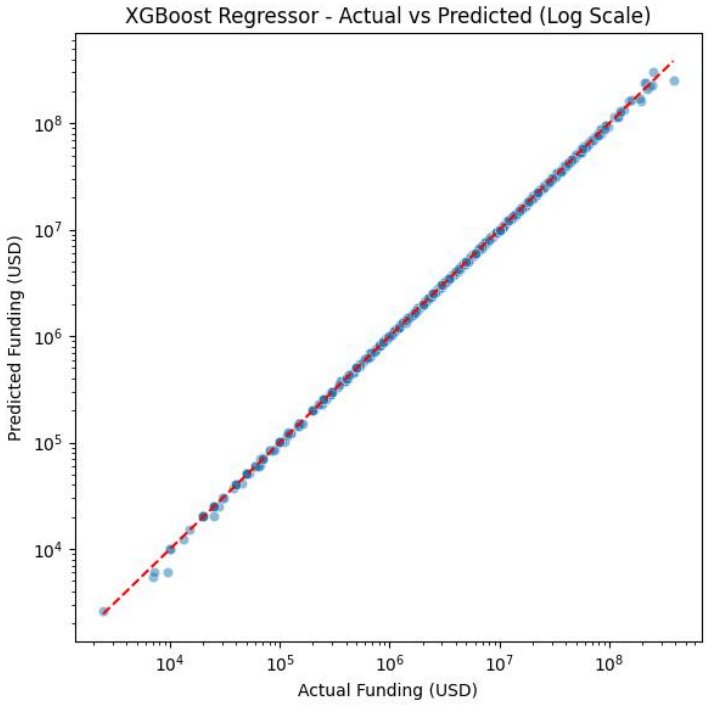
4.2.1. Giới thiệu về mô hình hồi quy

Mục tiêu của nghiên cứu là xây dựng và so sánh hiệu suất của các mô hình hồi quy trong việc dự đoán biến mục tiêu liên tục. Các mô hình được đánh giá gồm Linear Regression, Ridge Regression, Random Forest Regressor và XGBoost Regressor. Chúng tôi sử dụng các chỉ số chuẩn như RMSE (Root Mean Squared Error), R^2 Score, và kết quả Cross-Validation (CV R^2) để xác định mô hình tối ưu.

4.2.2. Kết quả mô hình

Mô hình	RMSE	R^2 Score	CV R^2 Mean	CV R^2 Std
Linear Regression	1.42336	0.566098	0.563720	0.025218
Ridge Regression	1.411116	0.566848	0.564406	0.025248
Random Forest Regressor	0.259838	0.985313	0.980409	0.003784
XGBoost Regressor	0.033814	0.999751	0.999660	0.000225

4.2.2.1. Biểu đồ so sánh giá trị thực tế và giá trị dự đoán



Hình 4.2 Biểu so sánh giá trị thực tế và giá trị dự đoán

Biểu đồ "XGBoost Regressor - Actual vs Predicted (Log Scale)" so sánh giá trị thực tế (Actual Funding USD) và giá trị dự đoán (Predicted Funding USD) của tổng vốn đầu tư (funding_total_usd) trên thang logarit, sử dụng mô hình XGBoost Regressor

1. Đánh giá hiệu suất dự đoán của mô hình: Biểu đồ cho thấy các điểm dữ liệu phân bố khá sát với đường chéo (y = x), đặc biệt ở các khoảng giá trị từ 10^4 đến

10^9 USD. Điều này chỉ ra rằng mô hình XGBoost Regressor có khả năng dự đoán tổng vốn đầu tư khá chính xác, với sự tương quan tốt giữa giá trị thực tế và giá trị dự đoán, đặc biệt trên thang logarit.

2. Ý nghĩa với bài toán dự đoán vốn đầu tư: Kết quả từ biểu đồ xác nhận hiệu quả của XGBoost trong việc dự đoán funding_total_usd, so sánh với các mô hình khác như Logistic Regression, Random Forest, SVM và KNN. Sự phân bố sát đường chéo cho thấy mô hình đã học tốt các đặc trưng (features) như loại hình đầu tư, thị trường, và độ tuổi startup, góp phần nâng cao độ tin cậy trong dự đoán vốn đầu tư.
3. Liên hệ với bài toán phân loại trạng thái hoạt động: Dữ liệu dự đoán vốn đầu tư chính xác từ XGBoost có thể hỗ trợ gián tiếp bài toán phân loại trạng thái hoạt động (đang hoạt động, đã bị mua lại, ngừng hoạt động). Ví dụ, các startup có vốn dự đoán cao có thể có xu hướng bị mua lại ("Acquired"), trong khi vốn thấp có thể liên quan đến việc đóng cửa ("Closed"). Điều này giúp cải thiện tính toàn diện của phân tích trong đề tài.
4. Hỗ trợ quyết định chiến lược cho nhà đầu tư: Biểu đồ cung cấp bằng chứng về độ tin cậy của mô hình XGBoost trong việc dự đoán vốn đầu tư, giúp nhà đầu tư đánh giá tiềm năng gọi vốn của startup một cách chính xác hơn. Các dự đoán sát với thực tế cho phép nhà đầu tư tối ưu hóa chiến lược đầu tư, tập trung vào các startup có tiềm năng nhận vốn cao, phù hợp với mục tiêu hỗ trợ quyết định chiến lược của đề tài.

Tóm lại, biểu đồ khẳng định hiệu suất của mô hình XGBoost Regressor trong việc dự đoán tổng vốn đầu tư, cung cấp cơ sở dữ liệu đáng tin cậy để hỗ trợ cả bài toán dự đoán và phân loại trạng thái hoạt động. Kết quả này góp phần nâng cao giá trị thực tiễn của đề tài, hỗ trợ nhà đầu tư đưa ra các quyết định chiến lược hiệu quả dựa trên dự đoán chính xác.

4.2.3. Phân tích kết quả

- Linear Regression

Mô hình tuyến tính cơ bản này cho thấy khả năng giải thích dữ liệu ở mức trung bình với $R^2 = 0.566$ và RMSE = 1.41. Kết quả từ cross-validation cũng phản ánh sự ổn định tương đối (CV R^2 Mean = 0.564, Std = 0.025). Tuy nhiên, mô hình này không đủ sức biểu diễn các quan hệ phi tuyến phức tạp trong dữ liệu thực tế.

- Ridge Regression

Là phiên bản mở rộng của Linear Regression có thêm thành phần điều chuẩn (regularization), Ridge mang lại kết quả tương đương với Linear Regression. RMSE (1.411) và R^2 (0.567) chỉ cải thiện nhẹ. Việc bổ sung regularization giúp kiểm soát overfitting tốt hơn, nhưng trong trường hợp này, mô hình vẫn bị giới hạn bởi giả định tuyến tính.

- Random Forest Regressor

Mô hình học máy dựa trên cây quyết định tổ hợp này thể hiện hiệu suất vượt trội hơn hẳn hai mô hình tuyến tính. Với RMSE giảm xuống còn 0.260 và R^2 tăng mạnh lên 0.985, Random Forest đã nắm bắt tốt các mối quan hệ phi tuyến trong dữ liệu. Đồng thời, mô hình có độ ổn định cao (CV R^2 Mean = 0.980, Std = 0.0038), cho thấy khả năng tổng quát hóa tốt.

- •XGBoost Regressor

Đây là mô hình có hiệu suất cao nhất trên toàn bộ các chỉ số. Với RMSE cực thấp (0.0338) và R^2 gần như tuyệt đối (0.9998), XGBoost thể hiện khả năng dự đoán chính xác gần như tuyệt đối. Hơn nữa, kết quả cross-validation rất ổn định (CV R^2 Mean = 0.9997, Std = 0.0002), khẳng định tính đáng tin cậy và tổng quát hóa cao của mô hình này. XGBoost kết hợp các ưu điểm của boosting và regularization, giúp kiểm soát overfitting hiệu quả.

4.2.4. Kết luận chi tiết: Tại sao XGBoost là mô hình tốt nhất

XGBoost Regressor được xác định là mô hình tốt nhất vì các lý do sau:

1. Độ chính xác cực cao:

RMSE rất nhỏ (0.0338) đồng nghĩa với sai số dự đoán gần như không đáng kể. Điều này cho thấy mô hình mô tả được gần như toàn bộ phân phối của biến mục tiêu.

2. Khả năng giải thích phương sai vượt trội:

Với R^2 đạt 0.999751, mô hình giải thích được gần như toàn bộ biến thiên trong dữ liệu. CV R^2 trung bình cũng đạt 0.999660, phản ánh hiệu quả học tốt trên tập huấn luyện và các tập kiểm tra trong quá trình Cross-Validation.

3. Tính ổn định và độ tin cậy cao:

Độ lệch chuẩn nhỏ (CV R^2 Std = 0.000225) chứng tỏ mô hình ổn định và ít biến động trong các lần lặp Cross-Validation, điều rất quan trọng khi áp dụng vào thực tế.

4. Tối ưu hóa hiệu quả và kiểm soát overfitting tốt:

XGBoost áp dụng nhiều kỹ thuật như regularization, shrinkage, và early stopping để kiểm soát overfitting, đồng thời tận dụng gradient boosting để cải thiện dự đoán theo từng bước huấn luyện.

XGBoost Regressor là lựa chọn tối ưu cho bài toán hồi quy hiện tại, nhờ vào khả năng mô phỏng chính xác dữ liệu, hiệu suất vượt trội và tính ổn định cao. Đây là mô hình nên được ưu tiên sử dụng trong triển khai thực tế.

CHƯƠNG 5. KẾT LUẬN, ƯU ĐIỂM, NHƯỢC ĐIỂM, ĐỊNH HƯỚNG PHÁT TRIỂN

5.1. KẾT LUẬN

Qua quá trình thử nghiệm và đánh giá bốn mô hình hồi quy có thể rút ra kết luận như sau:

Hai mô hình tuyến tính (Linear và Ridge Regression) có hiệu suất ở mức trung bình, phù hợp với các bài toán đơn giản hoặc khi cần giải thích mô hình rõ ràng.

XGBoost Regressor là mô hình có hiệu suất vượt trội nhất trên tất cả các chỉ số, thể hiện độ chính xác cao, sai số thấp và tính ổn định tốt. Đây là lựa chọn tối ưu cho bài toán dự đoán hiện tại.

5.2. ƯU ĐIỂM VÀ NHƯỢC ĐIỂM CỦA TỪNG MÔ HÌNH

Mô hình	Ưu điểm	Nhược điểm
Linear Regression	Dễ hiểu, dễ triển khai, tốc độ huấn luyện nhanh	Hiệu suất thấp, không xử lý được mối quan hệ phi tuyến
Ridge Regression	Giảm overfitting nhờ regularization, cải thiện nhẹ so với Linear	Hiệu suất vẫn còn thấp, phụ thuộc vào giả định tuyến tính
Random Forest	Hiệu suất cao, chống overfitting tốt, xử lý được quan hệ phi tuyến	Kém giải thích, chi phí tính toán cao hơn, có thể chậm với dữ liệu lớn
XGBoost Regression	Hiệu suất rất cao, ổn định, kiểm soát overfitting tốt, tối ưu hóa mạnh	Cấu hình phức tạp, cần tinh chỉnh siêu tham số, thời gian huấn luyện lâu hơn

5.2.1. Hướng phát triển

Để nâng cao chất lượng mô hình và khả năng ứng dụng trong thực tế, một số hướng phát triển đề xuất bao gồm:

1. Tối ưu siêu tham số tự động:
   Áp dụng các kỹ thuật tối ưu hóa siêu tham số như Grid Search, Random Search hoặc Bayesian Optimization để cải thiện thêm hiệu suất của XGBoost và Random Forest.
2. Feature Engineering:
   Thực hiện tạo thêm các đặc trưng mới, chọn lọc đặc trưng (feature selection) và chuẩn hóa dữ liệu hợp lý để giúp mô hình học tốt hơn.
3. Triển khai mô hình trong thực tế:
   Kết hợp mô hình XGBoost vào các hệ thống dự đoán, đồng thời xây dựng pipeline xử lý dữ liệu tự động và đánh giá hiệu suất theo thời gian thực.
4. Thử nghiệm thêm mô hình hiện đại:
   Khám phá các mô hình học sâu (deep learning) như Neural Networks, hoặc các biến thể tiên tiến của boosting như LightGBM, CatBoost để đối chiếu thêm.
5. Giải thích mô hình (Model Interpretability):
   Dù XGBoost hiệu quả, nhưng cần áp dụng thêm các kỹ thuật như SHAP hoặc LIME để giải thích và minh bạch hóa quyết định của mô hình, đặc biệt trong các ứng dụng cần tuân thủ đạo đức và pháp lý.

TÀI LIỆU THAM KHẢO

[1] AiCandy, “Thuật toán Random Forest: Giải thích chi tiết và ứng dụng,” [Trực tuyến]. Available: https://aicandy.vn/thuat-toan-random-forest-giai-thich-chi-tiet-va-ung-dung/. [Đã truy cập 1 5 2025].

[2] G. L. E. Team, “Hyperparameter Tuning with GridSearchCV,” [Trực tuyến]. Available: https://www.mygreatlearning.com/blog/gridsearchcv/. [Đã truy cập 25 4 2025].

[3] T. Vu, “Tại sao cần xây dựng pipeline,” [Trực tuyến]. Available: https://machinelearningcoban.com/tabml_book/ch_intro/why_pipeline.html. [Đã truy cập 16 4 2025].

[4] P. D. Khanh, “Sklearn Pipeline,” [Trực tuyến]. Available: https://phamdinhkhanh.github.io/deepai-book/ch_appendix/index_pipeline.html. [Đã truy cập 20 4 2025].

[5] Interdata, “XGBoost là gì? Cấu trúc, Tính năng & Ứng dụng trong học máy,” [Trực tuyến]. Available: https://interdata.vn/blog/xgboost-la-gi/. [Đã truy cập 20 5 2025].