TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO BÀI TẬP LỚN

HỌC PHẦN: HỌC MÁY

ĐỀ TÀI: XÂY DỰNG MÔ HÌNH DỰ BÁO RỦI RO TÀI CHÍNH DỰA TRÊN RANDOM FOREST

GVHD: TS TRẦN HÙNG CƯỜNG

Nhóm: 08

Lớp: 20232IT6047004

Sinh viên thực hiện: Vũ Xuân Điệp - 2021605707

Dương Việt Anh - 2021606084

Vũ Đình Chiểu - 2021605444

Ngô Thế Duy - 2021605852

Đỗ Thị Khánh Trinh -2021601095

Hà Nội – Năm 2024

MỤC LỤC

DANH MỤC HÌNH ẢNH

LỜI CẢM ƠN

Để hoàn thành được bài tập lớn này, nhóm chúng em xin được bày tỏ sự tri ân và chân thành cảm ơn giảng viên TS. Trần Hùng Cường người trực tiếp hướng dẫn, chỉ bảo nhóm chúng em trong suốt quá trình học tập và nghiên cứu để hoàn thành bài tập lớn này. Nhờ có sự giúp đỡ của thầy nhóm chúng em đã có thể hoàn thành bài tập một cách tốt nhất và đạt được những kết quả khả quan.

Trong quá trình nghiên cứu và làm báo cáo do năng lực, kiến thức còn hạn hẹp nên không tránh khỏi những thiếu sót. Nhóm chúng em kính mong nhận được sự thông cảm và những ý kiến đóng góp của quý thầy cô và các bạn.

Một lần nữa, em xin chân thành cảm ơn thầy/cô vì sự hỗ trợ và đóng góp của mình trong quá trình thực hiện bài tập lớn này.

Nhóm em xin chân thành cảm ơn!

LỜI MỞ ĐẦU

1. Lý do chọn đề tài

Trong những năm gần đây, việc ứng dụng các thuật toán học máy như Random Forest vào dự báo rủi ro tài chính đã trở nên ngày càng phổ biến. Việc xây dựng các mô hình dự báo chính xác các rủi ro tài chính là rất quan trọng, giúp các nhà đầu tư, doanh nghiệp và cả các cơ quan quản lý có thể đưa ra các quyết định đầu tư và quản lý rủi ro hiệu quả hơn.

Random Forest là một trong những thuật toán học máy phổ biến và mạnh mẽ, với khả năng xử lý dữ liệu phức tạp, nhiều chiều và khả năng dự báo chính xác cao. Vì vậy, việc nghiên cứu ứng dụng Random Forest vào dự báo rủi ro tài chính là rất hữu ích và có tiềm năng ứng dụng thực tế.

- Tính cấp thiết và tầm quan trọng của vấn đề: Dự báo và quản lý rủi ro tài chính là vấn đề then chốt đối với các tổ chức tài chính, doanh nghiệp và nhà đầu tư. Các mô hình dự báo chính xác các rủi ro tài chính sẽ giúp các chủ thể này đưa ra các quyết định đầu tư, quản lý rủi ro hiệu quả hơn, tránh được những tổn thất lớn do các biến động bất lợi trên thị trường tài chính.
- Tiềm năng ứng dụng của thuật toán học máy Random Forest: Random Forest là một trong những thuật toán học máy mạnh mẽ và hiệu quả trong các bài toán dự báo, phân loại. Việc nghiên cứu ứng dụng Random Forest vào dự báo rủi ro tài chính là một hướng nghiên cứu đầy hứa hẹn, có thể mang lại các mô hình dự báo chính xác và tin cậy hơn.
- Khả năng đóng góp cho lý thuyết và thực tiễn: Các mô hình dự báo được xây dựng có thể được ứng dụng trực tiếp trong hoạt động quản lý rủi ro của các tổ chức tài chính, doanh nghiệp.
- Tính khả thi của đề tài: Đề tài sử dụng dữ liệu công khai, các công cụ phân tích dữ liệu và thuật toán học máy phổ biến, do đó có tính khả thi cao và có thể được thực hiện trong khuôn khổ của một nghiên cứu.

Với những lý do trên, đề tài "Ứng dụng thuật toán Random Forest trong dự báo rủi ro tài chính" được lựa chọn là phù hợp và có ý nghĩa quan trọng về mặt lý thuyết cũng như thực tiễn.

2. Mục tiêu đề tài:

- Xây dựng mô hình dự báo rủi ro tài chính dựa trên thuật toán Random Forest.
- Đánh giá hiệu quả của mô hình dự báo dựa trên các chỉ số đánh giá mô hình.
- Đề xuất các giải pháp cải thiện hiệu suất của mô hình dự báo.

3. Phương pháp nghiên cứu:

- Thu thập và tiền xử lý dữ liệu về các chỉ số tài chính, kinh tế vĩ mô liên quan đến rủi ro tài chính.
- Xây dựng mô hình dự báo rủi ro tài chính dựa trên thuật toán Random Forest.
- Đánh giá hiệu suất của mô hình dự báo bằng các chỉ số như độ chính xác, độ nhạy, độ đặc hiệu, F1-score, v.v.
- Phân tích và đề xuất các giải pháp cải thiện hiệu suất của mô hình dự báo.

4. Đối tượng và phạm vi nghiên cứu:

- Đối tượng nghiên cứu: Mô hình dự báo rủi ro tài chính dựa trên thuật toán Random Forest.
- Phạm vi nghiên cứu: Áp dụng mô hình dự báo rủi ro tài chính cho các doanh nghiệp, tổ chức tài chính tại Việt Nam và thế giới.

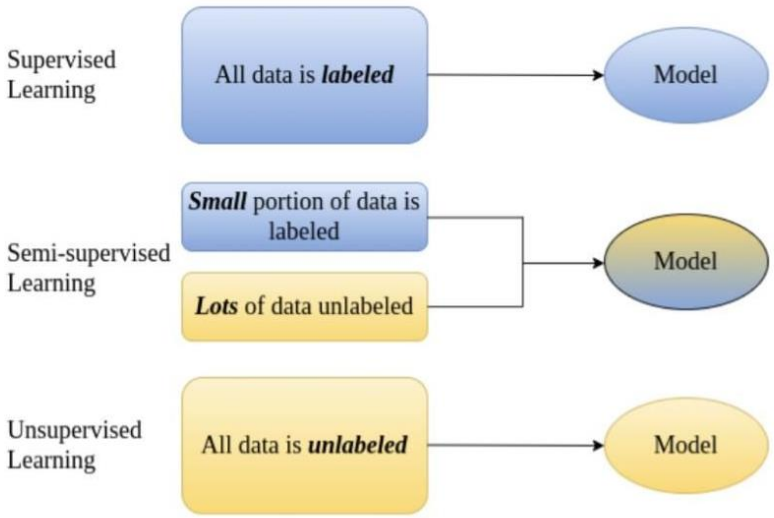
Chương 1. TỔNG QUAN VỀ HỌC MÁY

1.1. Khái niệm học máy

Machine Learning (hay học máy) là một lĩnh vực của trí tuệ nhân tạo (AI) mà nghiên cứu cách để máy tính học hỏi từ dữ liệu mà không cần được lập trình một cách cụ thể. Thay vì chỉ đơn giản là thực hiện các chỉ thị được lập trình trước đó, máy tính được lập trình để tìm ra các mô hình và kết luận từ dữ liệu đầu vào.

Machine Learning được áp dụng rộng rãi trong nhiều lĩnh vực khác nhau, bao gồm kinh doanh, y học, khoa học dữ liệu, thị giác máy tính, xử lý ngôn ngữ tự nhiên, và robot học. Với sự phát triển của công nghệ và dữ liệu, Machine Learning đang trở thành một công cụ quan trọng để giải quyết các vấn đề và tạo ra giá trị mới trong nhiều lĩnh vực.

1.2. Một số phương pháp học máy



Hình 1.1. Một số phương pháp học máy

1.2.1. Supervised machine learning (Học máy có giám sát)

Supervised learning còn được gọi là học máy có giám sát, được định nghĩa bằng cách sử dụng các tập dữ liệu được gắn nhãn để huấn luyện các thuật toán phân loại dữ liệu hoặc dự đoán kết quả một cách chính xác.

Khi dữ liệu đầu vào được đưa vào mô hình, mô hình sẽ điều chỉnh trọng lượng của nó cho đến khi nó được lắp một cách thích hợp. Điều này xảy ra như một phần của quá trình xác nhận chéo để đảm bảo rằng mô hình tránh trang bị quá nhiều hoặc trang bị thiếu thông tin.

Supervised machine learning giúp các tổ chức giải quyết nhiều vấn đề trong thế giới thực trên quy mô lớn, chẳng hạn như phân loại thư rác trong một thư mục riêng biệt từ hộp thư đến của bạn.

1.2.2. Unsupervised machine learning (Học máy không giám sát)

Unsupervised machine learning là phương pháp sử dụng các thuật toán máy học để phân tích và phân cụm các tập dữ liệu không được gắn nhãn.

Không cần sự can thiệp của con người, các thuật toán này có thể phát hiện ra các mẫu hoặc nhóm dữ liệu ẩn. Khả năng phát hiện ra những điểm tương đồng và khác biệt trong thông tin của phương pháp này khiến nó trở nên lý tưởng cho việc phân tích dữ liệu khám phá, chiến lược bán chéo (cross-sell), phân khúc khách hàng cũng như nhận dạng hình ảnh và mẫu.

Unsupervised machine learning cũng được sử dụng để giảm số lượng các tính năng trong một mô hình thông qua quá trình giảm kích thước. Phân tích thành phần chính (PCA) và phân tích giá trị đơn lẻ (SVD) là hai cách tiếp cận phổ biến cho việc này.

Các thuật toán khác được sử dụng trong học tập không giám sát bao gồm: k-means clustering, neural networks, và probabilistic clustering methods.

Một số phương pháp được sử dụng trong Supervised machine learning bao gồm: logistic regression, neural networks, linear regression, naive bayes, random forest, và support vector machine (SVM).

1.2.3. Semi-supervised learning

Semi-supervised learning cung cấp một phương pháp hiệu quả giữa học tập có giám sát và không giám sát. Trong quá trình đào tạo, nó sử dụng một tập dữ liệu có nhãn nhỏ hơn để hướng dẫn phân loại và trích xuất tính năng từ một tập dữ liệu lớn hơn, không được gắn nhãn.

Phương pháp Semi-supervised learning có thể giải quyết vấn đề không có đủ dữ liệu được gắn nhãn cho thuật toán học có giám sát. Nó cũng hữu ích nếu quá tốn kém để gắn nhãn đủ dữ liệu.

1.3. Ứng dụng của học máy

Một số ứng dụng thực tế của Machine Learning như:

Speech recognition: Dùng để nhận dạng giọng nói tự động (ASR), nhận dạng giọng nói máy tính hoặc chuyển giọng nói thành văn bản. Đây là một khả năng sử dụng xử lý ngôn ngữ tự nhiên (NLP) để dịch giọng nói của con người sang định dạng viết.

Customer service: Chatbots trực tuyến đang thay thế các tác nhân con người trong hành trình của khách hàng, thay đổi cách chúng ta nghĩ về sự tương tác của khách hàng trên website và nền tảng xã hội.

Computer vision: Công nghệ AI này cho phép máy tính lấy thông tin có ý nghĩa từ video, hình ảnh kỹ thuật số và các đầu vào trực quan khác, sau đó thực thi hành động thích hợp.

Recommendation engines: Sử dụng dữ liệu hành vi tiêu dùng trong quá khứ, các thuật toán AI learning có thể giúp khám phá các xu hướng dữ liệu có thể được sử dụng để phát triển các chiến lược cross-sell hiệu quả hơn.

Automated stock trading: Được thiết kế để tối ưu hóa danh mục đầu tư chứng khoán, các nền tảng giao dịch tần suất cao do AI điều khiển để hàng triệu giao dịch mỗi ngày mà không cần đến sự can thiệp của con người.

Fraud detection: Các ngân hàng và các tổ chức tài chính có thể sử dụng máy học để phát hiện các giao dịch đáng ngờ.

1.4. Những thách thức trong học máy

Chất lượng dữ liệu: Một trong những thách thức lớn nhất là có được một tập dữ liệu đủ lớn, đầy đủ và chất lượng cao. Dữ liệu thường bị thiếu, lộn xộn hoặc có nhiều lỗi, ảnh hưởng đáng kể đến hiệu suất của mô hình.

Overfitting và underfitting: Mô hình có thể bị overfitting, nghĩa là quá phù hợp với tập dữ liệu huấn luyện nhưng lại kém hiệu quả trên dữ liệu mới. Ngược lại, mô hình cũng có thể bị underfitting do không đủ phức tạp để mô tả được mối quan hệ trong dữ liệu.

Tính khả giải thích: Nhiều mô hình học máy hoạt động như "hộp đen", khó giải thích được cách chúng đưa ra quyết định. Điều này gây khó khăn khi triển khai vào các ứng dụng có yêu cầu về tính minh bạch.

Bias và Fairness: Mô hình học máy có thể mang theo các loại bias (thiên vị) như giới tính, chủng tộc, tuổi tác, v.v. dẫn đến các quyết định không công bằng. Vì vậy, cần đảm bảo tính công bằng của mô hình.

Dữ liệu động và concept drift: Trong thực tế, dữ liệu và mối quan hệ thường thay đổi theo thời gian (concept drift). Điều này đòi hỏi mô hình phải liên tục được cập nhật và tinh chỉnh.

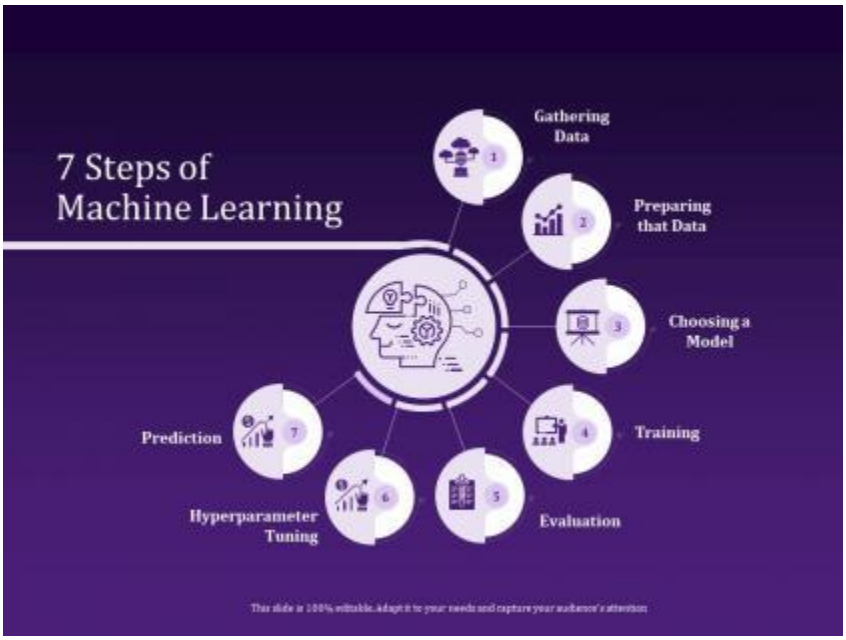
Tài nguyên tính toán: Huấn luyện và triển khai các mô hình học máy, đặc biệt là mô hình phức tạp, thường đòi hỏi nhiều tài nguyên tính toán như bộ nhớ, CPU, GPU.

Bảo mật và riêng tư: Khi sử dụng dữ liệu cá nhân để huấn luyện mô hình, cần đảm bảo an ninh và bảo mật thông tin, đồng thời tuân thủ các quy định về quyền riêng tư.

Các thách thức này đòi hỏi các kỹ sư và nhà khoa học dữ liệu phải có kiến thức và kỹ năng chuyên sâu để có thể thiết kế, huấn luyện và triển khai các mô hình học máy hiệu quả và tin cậy.

1.5. Các bước cơ bản để xây dựng mô hình học máy

Các bước xây dựng mô hình học máy



Hình 1.2. Minh họa các bước xây dựng mô hình học máy

- Bước 1: Thu thập dữ liệu

Việc đầu tiên trong quá trình xây dựng mô hình học máy là thu thập dữ liệu phù hợp. Bạn cần tìm kiếm và thu thập một tập dữ liệu đầy đủ, chính xác và đại diện cho vấn đề muốn giải quyết. Chất lượng của dữ liệu có ảnh hưởng trực tiếp đến hiệu suất của mô hình.

- Bước 2: Chuẩn bị dữ liệu

Sau khi thu thập dữ liệu, bước tiếp theo là chuẩn bị dữ liệu. Bạn cần làm sạch dữ liệu bằng cách xử lý các giá trị thiếu, loại bỏ các điểm ngoại lai, và chuẩn hóa dữ

liệu để đảm bảo chúng ở cùng một thang đo. Cuối cùng, chia dữ liệu thành tập huấn luyện và tập kiểm tra.

- Bước 3: Chọn mô hình

Sau khi dữ liệu đã sẵn sàng, bước tiếp theo là chọn mô hình học máy phù hợp với bài toán. Bạn cần xem xét các mô hình khác nhau như hồi quy, phân loại, cụm, v.v. và lựa chọn mô hình tối ưu dựa trên đặc điểm của dữ liệu và yêu cầu của bài toán.

- Bước 4: Huấn luyện mô hình

Khi đã chọn được mô hình phù hợp, bước tiếp theo là huấn luyện mô hình. Sử dụng tập dữ liệu huấn luyện, bạn sẽ điều chỉnh các tham số của mô hình bằng cách sử dụng các thuật toán huấn luyện thích hợp.

- Bước 5: Đánh giá mô hình

Sau khi huấn luyện mô hình, bạn cần đánh giá hiệu suất của nó trên tập dữ liệu kiểm tra. Tính toán các chỉ số đánh giá như độ chính xác, độ nhạy, độ đặc hiệu, v.v. để xác định xem mô hình có đạt yêu cầu hay không.

- Bước 6: Điều chỉnh tham số

Nếu kết quả chưa đạt yêu cầu, bạn cần tiếp tục điều chỉnh các siêu tham số của mô hình. Thử nghiệm các giá trị siêu tham số khác nhau để tìm ra cấu hình tối ưu.

- Bước 7: Đưa ra dự đoán

Khi mô hình đã được huấn luyện và điều chỉnh, bạn có thể sử dụng nó để đưa ra dự đoán trên dữ liệu mới. Cuối cùng, triển khai mô hình vào thực tế để giải quyết vấn đề.

1.6. Đánh giá hiệu năng mô hình phân lớp và mô hình dự báo

Mô hình phân lớp:

- Accuracy (Độ chính xác): Tỷ lệ giữa số lượng dự đoán đúng và tổng số dự đoán.

- Precision (Độ chính xác trung bình ở mỗi lớp):

Tỷ lệ giữa số lượng dự đoán đúng trong số các dự đoán là dương tính.

Precision = TP / (TP + FP), trong đó TP là số lượng True Positives và FP là số lượng False Positives.

- Recall (Độ thu hồi trung bình ở mỗi lớp):

Tỷ lệ giữa số lượng dự đoán đúng trong số các trường hợp dương tính thực sự.

Recall = TP / (TP + FN), trong đó FN là số lượng False Negatives.

- F1 Score:

Trung bình điều hòa của Precision và Recall.

F1 Score = 2 * (Precision * Recall) / (Precision + Recall).

- ROC-AUC (Receiver Operating Characteristic - Area Under Curve):

Đánh giá khả năng phân biệt giữa các lớp của mô hình. AUC cao hơn thể hiện mô hình phân loại tốt hơn.

- Confusion Matrix (Ma trận nhầm lẫn):

Biểu diễn số lượng True Positives (TP), True Negatives (TN), False Positives (FP), và False Negatives (FN).

Mô hình dự đoán:

- Mean Absolute Error- MAE (sai số tuyệt đối trung bình):

Trung bình của các giá trị tuyệt đối của sai số giữa giá trị dự đoán và giá trị thực tế.

$$MAE = (1/n) * \Sigma |y_i - \hat{y}_i|$$

- Mean Squared Error-MSE(Sai số bình phương trung bình):

Trung bình của các giá trị bình phương của sai số giữa giá trị dự đoán và giá trị thực tế.

$$MSE = (1/n) * \Sigma (y_i - \hat{y}_i)^2$$

- Root Mean Squared Error (RMSE):

Căn bậc hai của MSE, đo lường sai số trung bình của mô hình

$$RMSE = sqrt(MSE)$$

- R-squared (R^2) or Coefficient of Determination(hệ số xác định):

Đo lường tỷ lệ phương sai của biến phụ thuộc được giải thích bởi biến độc lập trong mô hình.

$R^2 = 1 - (SS\_res / SS\_tot)$, trong đó SS_res là tổng bình phương của sai số còn lại, và SS_tot là tổng bình phương của sai số của tổng thể.

- Adjusted R-squared:

Phiên bản điều chỉnh của R^2 để tránh trường hợp thêm biến không hữu ích vào mô hình.

Adjusted $R^2 = 1 - [(1 - R^2) * (n - 1) / (n - p - 1)]$, trong đó n là số lượng quan sát và p là số lượng biến độc lập.

Chương 2. THUẬT TOÁN RANDOM FOREST

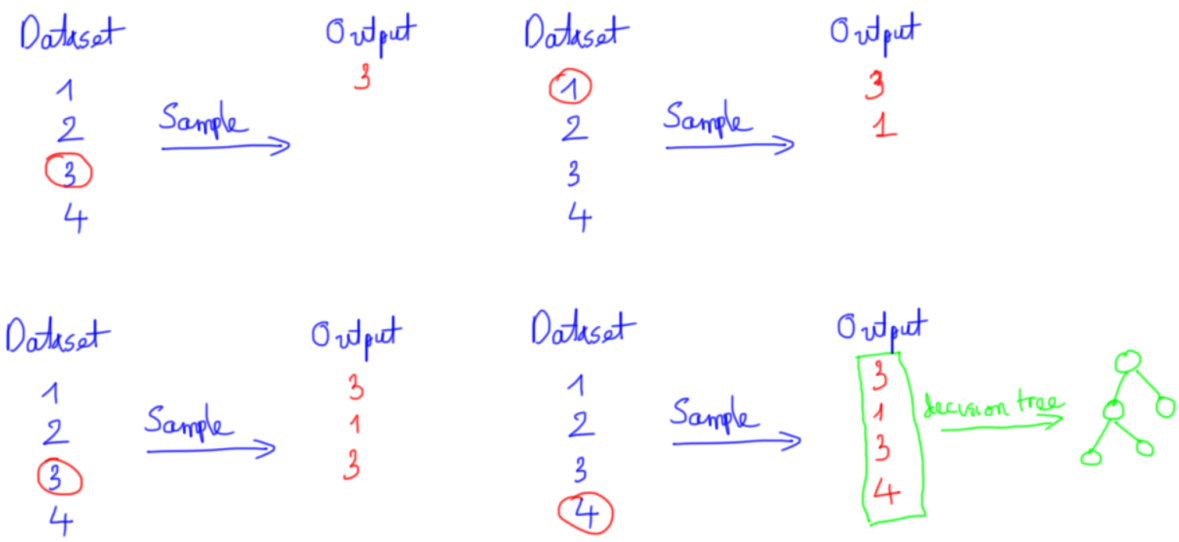
2.1. Khái niệm thuật toán Random Forest

Thuật toán Random Forest là một thuật toán học có giám sát được sử dụng để phân loại hoặc dự đoán giá trị của một biến mục tiêu dựa trên các thuộc tính đầu vào. Nó là một phương pháp kết hợp nhiều cây quyết định để tạo ra một mô hình phân loại chính xác hơn.

2.2. Nguyên tắc hoạt động thuật toán Random Forest

Mô hình rừng cây được huấn luyện dựa trên sự phối hợp giữa luật kết hợp (ensembling) và quá trình lấy mẫu tái lặp (bootstrapping). Cụ thể thuật toán này tạo ra nhiều cây quyết định mà mỗi cây quyết định được huấn luyện dựa trên nhiều mẫu con khác nhau và kết quả dự báo là bầu cử (voting) từ toàn bộ những cây quyết định. Như vậy một kết quả dự báo được tổng hợp từ nhiều mô hình nên kết quả của chúng sẽ không bị chệch. Đồng thời kết hợp kết quả dự báo từ nhiều mô hình sẽ có phương sai nhỏ hơn so với chỉ một mô hình.

Lấy ngẫu nhiên n dữ liệu từ bộ dữ liệu với kỹ thuật bootstrapping, hay còn gọi là random sampling with replacement. Tức khi mình sample được 1 dữ liệu thì mình không bỏ dữ liệu đấy ra mà vẫn giữ lại trong tập dữ liệu ban đầu, rồi tiếp tục sample cho tới khi sample đủ n dữ liệu. Khi dùng kĩ thuật này thì tập n dữ liệu mới của mình có thể có những dữ liệu bị trùng nhau.

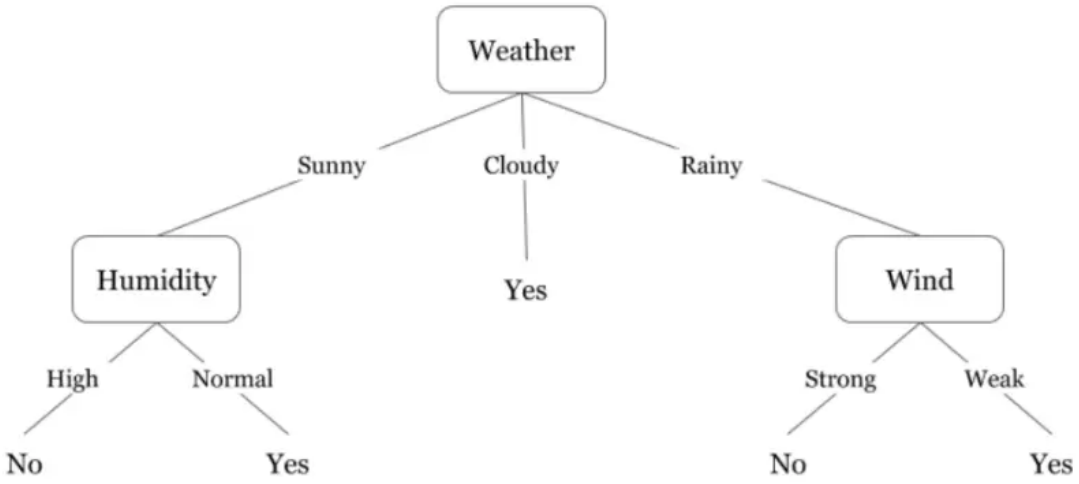


Hình 2.1. Minh họa cách tạo ra các dataset ngẫu nhiên

Sau khi sample được n dữ liệu từ bước 1 thì mình chọn ngẫu nhiên ở k thuộc tính (k < n). Giờ mình được bộ dữ liệu mới gồm n dữ liệu và mỗi dữ liệu có k thuộc tính. Dùng thuật toán Decision Tree (ID3) để xây dựng cây quyết định với bộ dữ liệu.

Mô hình rừng cây là sự kết hợp của nhiều cây quyết định và để hiểu được chúng ta cần hiểu thuật toán cây quyết định (ID3)

Cây quyết định là một mô hình học máy có giám sát, có thể được áp dụng vào cả hai bài toán phân lớp và hồi quy. Cụ thể, cây quyết định là một cấu trúc giống như lưu đồ, trong đó mỗi nút bên trong đại diện cho một "thử nghiệm" trên một thuộc tính (ví dụ: liệu một lần lật đồng xu xuất hiện mặt ngửa hay sấp), mỗi nhánh đại diện cho kết quả của thử nghiệm và mỗi nút lá đại diện cho một nhãn lớp (quyết định được thực hiện sau khi tính toán tất cả các thuộc tính). Các đường đi từ gốc đến lá đại diện cho các quy tắc phân loại.



Hình 2.2. Minh họa cây quyết định

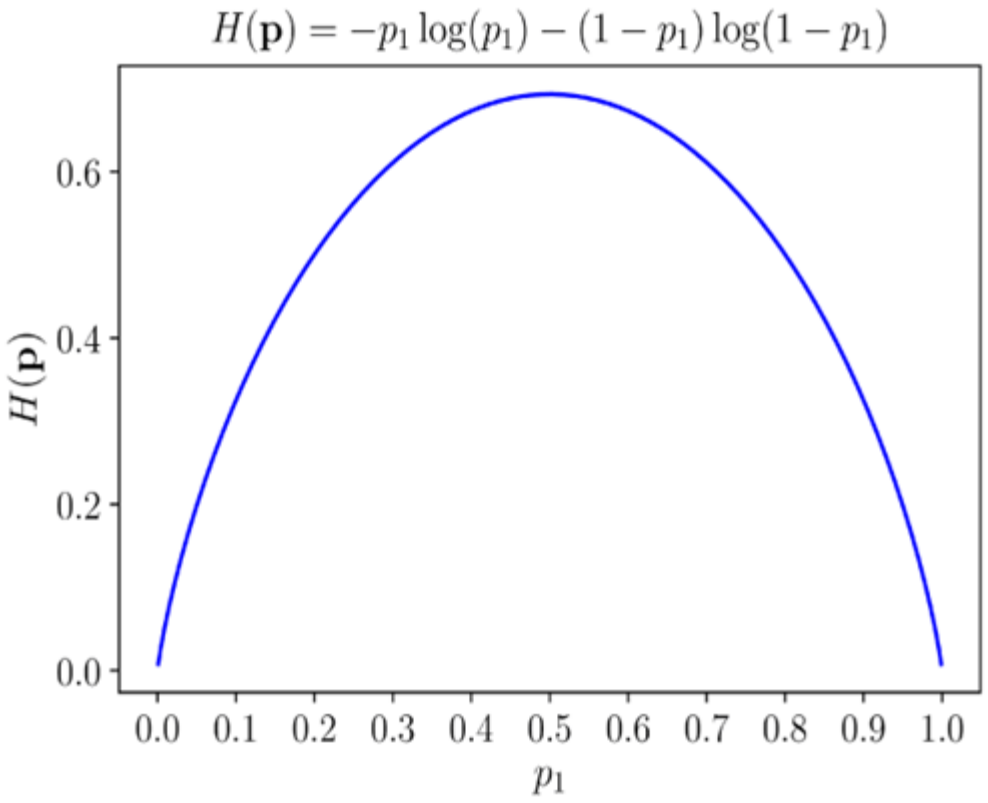
Thuật toán ID3 là một thuật toán nhằm xây dựng cây quyết định được áp dụng cho các bài toán phân lớp mà tất cả các thuộc tính đều ở dạng dữ liệu phân loại (các thông tin có đặc điểm giống nhau được nhóm lại, VD: (mưa, nắng) hay (xanh, đỏ…). Trong ID3, ta cần xác định thứ tự của thuộc tính cần được xem xét tại mỗi bước. Với các bài toán có nhiều thuộc tính và mỗi thuộc tính có nhiều giá trị khác nhau, việc tìm được nghiệm tối ưu thường là không khả thi. Thay vào đó, một phương pháp đơn giản thường được sử dụng là tại mỗi bước, một thuộc tính tốt nhất sẽ được chọn ra dựa trên một tiêu chuẩn định trước nào đó. Với mỗi thuộc tính được chọn, ta chia dữ liệu vào các "child node" tương ứng với các giá trị của thuộc tính đó rồi tiếp tục áp dụng phương pháp này cho mỗi đỉnh con. Việc chọn ra thuộc tính tốt nhất ở mỗi bước như thế này được gọi là cách chọn greedy (tham lam). Cách chọn này có thể không phải là tối ưu, nhưng trực giác cho chúng ta thấy rằng cách làm này sẽ gần với cách làm tối ưu. Ngoài ra, cách làm này khiến cho bài toán cần giải quyết trở nên đơn giản hơn. Sau mỗi câu hỏi để phân chia tại các đỉnh, dữ liệu được chia vào từng đỉnh con tương ứng với các câu trả lời cho câu hỏi đó. Câu hỏi ở đây chính là

một thuộc tính, câu trả lời chính là giá trị của thuộc tính đó. Để đánh giá chất lượng của một cách phân chia, chúng ta cần đi tìm một phép đo.

Trước hết, thế nào là một phép phân chia tốt? Bằng trực giác, một phép phân chia là tốt nhất nếu dữ liệu trong mỗi đỉnh con hoàn toàn thuộc vào một lớp khi đó đỉnh con này có thể được coi là một lá, tức ta không cần phân chia thêm nữa. Nếu dữ liệu trong các đỉnh con vẫn lẫn vào nhau theo tỉ lệ lớn, ta coi rằng phép phân chia đó chưa thực sự tốt. Từ nhận xét này, ta cần có một hàm số đo độ tinh khiết (purity), hoặc độ vẩn đục (impurity) của một phép phân chia. Hàm số này sẽ cho giá trị thấp nhất nếu dữ liệu trong mỗi đỉnh con nằm trong cùng một lớp tinh khiết nhất, và cho giá trị cao nếu mỗi đỉnh con có chứa dữ liệu thuộc nhiều lớp khác nhau. Một hàm số có các đặc điểm này và được dùng nhiều trong lý thuyết thông tin là hàm entropy. Cho một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau x_1, x_2,…x_n. Giả sử rằng xác suất để x nhận các giá trị này là$p_i = p(x = x_i)$ với $0 \leq p_i \leq x_i$, $\sum_{i=1}^{n} p_{i=1}$. Ký hiệu phân phối này là p=(p_1, p_2,…, p_n). Entropy của phân phối này được định nghĩa là:

$$H(p) = \sum_{i=1}^{n} p_i log(p_i)$$

Trong đó log là logarit tự nhiên (hoặc một số tài liệu dùng là logarit cơ số 2, nhưng giá trị của H(p) chỉ khác bằng cách nhân với một hằng số) và quy ước 0 logf_0 $f_0$$f_0$(0)=0.



Hình 2.3. Đồ thị hàm entropy với n = 2

Khi đó p tinh khiết nhất khi p_i là 1 trong 2 giá trị là 0 hoặc 1; và p bị vẩn đục khi p_i=0,5. Tổng quát lên với n > 2, hàm entropy đạt giá trị nhỏ nhất nếu có một giá trị p_i=1 và lớn nhất nếu tất cả p_i bằng nhau. Những tính chất này của hàm entropy khiến nó được sử dụng trong việc đo độ vẩn đục của một phép phân chia của ID3. Vì lý do này, ID3 còn được gọi là entropy-based decision tree.

Trong ID3, tổng có trọng số của entropy tại các lá sau khi xây dựng cây quyết định được coi là hàm mất mát của cây quyết định đó. Các trọng số ở đây tỉ lệ với số điểm dữ liệu được phân vào mỗi đỉnh. Công việc của ID3 là tìm các cách phân chia hợp lý (thứ tự chọn thuộc tính hợp lý) sao cho hàm mất mát cuối cùng đạt giá trị càng nhỏ càng tốt. Như đã đề cập, việc này đạt được bằng cách chọn ra thuộc tính sao cho nếu dùng thuộc tính đó để phân chia, entropy tại mỗi bước giảm đi một lượng lớn nhất. Bài toán xây dựng một cây quyết định bằng ID3 có thể chia thành

các bài toán nhỏ, trong mỗi bài toán, ta chỉ cần chọn ra thuộc tính giúp cho việc phân chia đạt kết quả tốt nhất. Mỗi bài toán nhỏ này tương ứng với việc phân chia dữ liệu trong một đỉnh không phải lá. Chúng ta sẽ xây dựng phương pháp tính toán dựa trên mỗi đỉnh này.

Xét một bài toán với C lớp khác nhau. Giả sử ta đang làm việc với một đỉnh không phải lá với các điểm dữ liệu tạo thành một tập S với số phần tử là |S|=N. Giả sử thêm rằng trong số N điểm dữ liệu này, N_c với c=1,2,…,C điểm thuộc vào lớp c. Xác suất để mỗi điểm dữ liệu rơi vào một lớp c được xấp xỉ bằng N_c/N (Ước lượng khả năng tối đa). Như vậy, giá trị entropy tại đỉnh này được tính bởi:

$$H(S) = -\sum_{c=1}^{C} \frac{N_C}{N} log\ log(\frac{N_C}{N})$$

Tiếp theo, giả sử thuộc tính được chọn là x. Dựa trên x, các điểm dữ liệu trong S được chia thành K đỉnh con: S_1,S_2,…,S_K với số điểm trong mỗi đỉnh con lần lượt là m_1,m_2,…m_K. Ta định nghĩa:

$$H(x,S) = -\sum_{k=1}^{K} \frac{M_k}{N} H(S_k)$$

Là tổng có trọng số entropy của mỗi đỉnh con được tính tương tự như H(S). Việc lấy trọng số này rất quan trọng vì các đỉnh thường có số lượng điểm khác nhau. Tiếp theo, ta định nghĩa Information gain dựa trên thuộc tính của x:

G(x,S) = H(S) - H(x,S)

Trong ID3, tại mỗi đỉnh, thuộc tính được chọn được xác định bởi

x* = argG(x,S) = argH(x,S)

Tức là thuộc tính khiến Information gain đạt giá trị lớn nhất.

Ngoài ra đối với ID3 hồi quy ta lại xây dựng cây quyết định dựa trên các thông số khác bao gồm phương sai, độ giảm phương sai, hệ số biến thiên.

Phương sai (Variance): Phương sai là một độ đo cho biết mức độ phân tán của dữ liệu. Khi xây dựng cây quyết định, chúng ta có thể lựa chọn các thuộc tính để chia nhánh sao cho phương sai của các nhóm con là nhỏ nhất. Điều này giúp đảm bảo rằng các điểm dữ liệu trong mỗi nhóm con tập trung gần nhau, tăng tính đồng đều và đồng nhất của nhóm.

Cho một tập dữ liệu D có n điểm dữ liệu, là giá trị của từng điểm và là trung bình cộng của n điểm phương sai được tính như sau:

$$\text{Variance}(D) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Độ giảm phương sai (Variance Reduction): Độ giảm phương sai đo lường sự giảm thiểu phương sai tổng cộng khi chia nhánh dữ liệu. Khi chọn thuộc tính để chia, chúng ta cố gắng chọn thuộc tính sao cho việc chia nhánh giảm phương sai nhiều nhất, tức là làm giảm độ biến động của dữ liệu.

Cho một thuộc tính A và một giá trị ngưỡng θ, độ giảm phương sai được tính bằng sự giảm thiểu phương sai từ việc chia nhánh dữ liệu:

$$\text{Variance Reduction}(D, A, \theta) = \text{Variance}(D) - \left(\frac{|D_1|}{|D|}\text{Variance}(D_1) + \frac{|D_2|}{|D|}\text{Variance}(D_2)\right)$$

Một vấn đề đặt ra trong quá trình xây cây quyết định là với các trường dữ liệu có kiểu số và có nhiều giá trị khác nhau thì ta sẽ thực hiện đánh giá như thế nào?. Việc này sẽ được thực hiện bằng cách tìm ngưỡng để chia trường dữ liệu ra làm hai trường hợp đưa về dạng giống với các trường dữ liệu có hữu hạn các giá trị khác

nhau. Để xác định ngưỡng của một trường dữ liệu giả sử là A ta sẽ thực hiện như sau:

- Bước 1: Sắp xếp các giá trị của A theo chiều tăng dần, trường mục tiêu được sắp xếp theo A. Điều này giúp trong việc dễ dàng lấy giá trị trung bình cộng của hai giá trị liền kề để tạo ra ngưỡng.
- Bước 2: Duyệt qua các ngưỡng. Một vòng lặp chạy qua từng giá trị trong dãy A đã sắp xếp để tạo ra các ngưỡng. Đối với mỗi ngưỡng, dữ liệu trong A lại được chia thành hai phần: một phần có giá trị nhỏ hơn hoặc bằng ngưỡng và một phần có giá trị lớn hơn ngưỡng. Đối với mỗi ngưỡng, MSE được tính toán trên trường mục tiêu cho việc phân chia dữ liệu. MSE là sự tổng của bình phương độ chênh lệch giữa giá trị thực tế và giá trị trung bình của mỗi phần. Công thức tính như sau:

$$MSE = \sum_{i=1}^{n1} (y_i - \underline{y1})^2 + \sum_{j=1}^{n2} (y_j - \underline{y2})^2$$

Trong đó:

- n1: là số giá trị của trường mục tiêu y ứng với phần nhỏ hơn hoặc bằng ngưỡng của A.

- $\underline{y1}$: là trung bình cộng của n1 giá trị của trường mục tiêu.

- n2: là số giá trị của trường mục tiêu y ứng với phần lớn hơn ngưỡng của A.

- $\underline{y2}$: là trung bình cộng của n2 giá trị của trường mục tiêu.

Sau khi tính xong, so sánh các MSE và lưu giữ ngưỡng tốt nhất. Ngưỡng tạo ra MSE thấp nhất được chọn làm ngưỡng tốt nhất. MSE thấp nhất thường là mục tiêu vì nó thể hiện sự giảm thiểu sự chênh lệch giữa giá trị thực tế và dự đoán.

Qua các bước này ta xác định được ngưỡng tốt nhất để phân chia dữ liệu trong A thành hai trường hợp nhỏ hơn hoặc bằng ngưỡng và lớn hơn ngưỡng. Từ đó giúp cho việc đánh giá A trở nên dễ dàng.

Điều kiện dừng của thuật toán ID3: Trong các thuật toán cây quyết định nói chung và ID3 nói riêng, nếu ta tiếp tục phân chia các đỉnh chưa tinh khiết, ta sẽ thu được một cây mà mọi điểm trong tập huấn luyện đều được dự đoán đúng (giả sử rằng không có hai đầu vào giống nhau nào cho đầu ra khác nhau). Khi đó, cây có thể sẽ rất phức tạp (nhiều đỉnh) với nhiều lá chỉ có một vài điểm dữ liệu. Như vậy, nhiều khả năng overfitting sẽ xảy ra. Để tránh overfitting, một trong số các phương pháp sau có thể được sử dụng. Tại một đỉnh, nếu một trong số các điều kiện sau đây xảy ra, ta không tiếp tục phân chia node đó và coi nó là một lá:

- Nếu đỉnh đang xét có entropy , bằng 0, tức mọi điểm ở đỉnh đều thuộc một lớp.

- Nếu đỉnh đó có số phần tử nhỏ hơn một ngưỡng nào đó. Trong trường hợp này, ta chấp nhận có một số điểm bị phân lớp sai để tránh overfitting. Lớp cho lá này có thể được xác định dựa trên lớp chiếm đa số trong đỉnh.

- Nếu khoảng cách từ đỉnh đó đến đỉnh gốc đạt tới một giá trị nào đó. Việc hạn chế chiều sâu của cây này làm giảm độ phức tạp của cây và phần nào giúp tránh overfitting.

- Nếu tổng số lá vượt quá một ngưỡng nào đó.

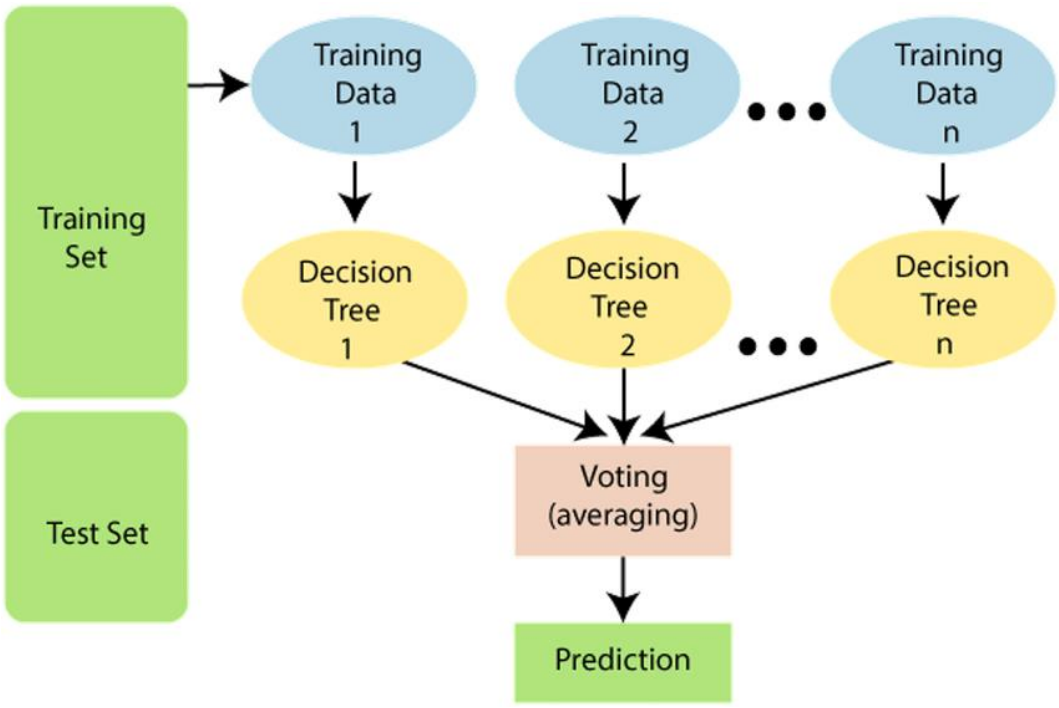
- Nếu việc phân chia đỉnh đó không làm giảm entropy quá nhiều (information gain nhỏ hơn một ngưỡng nào đó).

- Ngoài các phương pháp trên, một phương pháp phổ biến khác được sử dụng để tránh overfitting là pruning.

Như vậy, từ những thông tin và ví dụ ở trên ta đã hiểu được về thuật toán ID3 và mô hình cây quyết định. Về mặt kỹ thuật, RF là một phương pháp tổng hợp (dựa trên cách tiếp cận phân chia và chinh phục) của các cây quyết định được tạo ra trên một tập dữ liệu được chia ngẫu nhiên. Bộ sưu tập phân loại cây quyết định này còn được gọi là rừng. Cây quyết định riêng lẻ được tạo ra bằng cách sử dụng chỉ báo chọn thuộc tính như tăng thông tin, tỷ lệ tăng và chỉ số Gain (hoặc Gini) cho từng thuộc tính. Mỗi cây phụ thuộc vào một mẫu ngẫu nhiên độc lập. Trong bài toán phân loại, mỗi phiếu bầu chọn và lớp phổ biến nhất được chọn là kết quả cuối cùng. Trong trường hợp hồi quy, mức trung bình của tất cả các kết quả đầu ra của cây được coi là kết quả cuối cùng. Nó đơn giản và mạnh mẽ hơn so với các thuật toán phân loại phi tuyến tính khác.

2.3. Các bước thực hiện

- Chọn ngẫu nhiên một tập con các mẫu dữ liệu từ tập dữ liệu ban đầu.
- Xây dựng một cây quyết định trên tập con dữ liệu được chọn. Khi xây dựng cây, ta chỉ chọn ngẫu nhiên một số thuộc tính để xem xét khi tìm kiếm thuộc tính tốt nhất để chia tập dữ liệu.
- Lặp lại quá trình trên để xây dựng nhiều cây quyết định khác nhau.
- Khi có một mẫu dữ liệu mới cần phân loại, ta cho mỗi cây quyết định trả về kết quả phân loại của nó. Kết quả cuối cùng được tính toán bằng cách chọn lớp có số phiếu bầu (votes) cao nhất.



Hình 2.4. Các bước thực hiện thuật toán Random Forest

2.4. Ưu điểm và nhược điểm của thuật toán

Ưu điểm

- Tính ổn định: Random Forest có khả năng giảm thiểu overfitting, tức là mô hình không chỉ tốt trên tập huấn luyện mà còn tốt trên tập kiểm tra.
- Tính linh hoạt: Random Forest có thể được sử dụng cho cả bài toán phân loại và dự đoán.
- Tính khả diễn giải: Random Forest có thể cung cấp thông tin về sự quan trọng của các thuộc tính đầu vào trong việc phân loại.

Nhược điểm

- Tốc độ huấn luyện: Do phải xây dựng nhiều cây quyết định, việc huấn luyện Random Forest có thể mất nhiều thời gian hơn so với các thuật toán khác.
- Tính phức tạp: Random Forest có nhiều siêu tham số cần được điều chỉnh để đạt hiệu suất tốt nhất, điều này có thể khiến cho việc sử dụng thuật toán này trở nên phức tạp hơn so với các thuật toán khác

Chương 3. MÔ HÌNH DỰ BÁO RỦI RO TÀI CHÍNH DỰA TRÊN RANDOM FOREST

3.1. Giới thiệu

Trong bối cảnh tài chính phức tạp và đầy biến động ngày nay, việc dự đoán rủi ro tài chính trở thành một yếu tố then chốt giúp các tổ chức duy trì sự ổn định và phát triển bền vững. Mô hình dự báo rủi ro tài chính dựa trên Random Forest là một giải pháp công nghệ tiên tiến, kết hợp giữa sức mạnh của học máy và khả năng phân tích dữ liệu vượt trội để đưa ra các dự báo chính xác và kịp thời.

Random Forest, một trong những thuật toán học máy mạnh mẽ nhất, hoạt động bằng cách xây dựng một tập hợp các cây quyết định (decision trees) và tổng hợp kết quả của chúng. Điều này không chỉ cải thiện độ chính xác của dự báo mà còn giúp giảm thiểu nguy cơ quá khớp (overfitting), đảm bảo rằng mô hình hoạt động tốt trên cả dữ liệu huấn luyện và dữ liệu thực tế.

Mô hình này phân tích một loạt các yếu tố quan trọng như nghề nghiệp, tuổi, tình trạng hôn nhân, trình độ học vấn, tình trạng nợ, số dư tài khoản, và các yếu tố tài chính khác. Thông qua việc khai thác những mẫu và xu hướng tiềm ẩn trong dữ liệu, mô hình có thể dự đoán khả năng xảy ra rủi ro tài chính của cá nhân hoặc tổ chức một cách chính xác và tin cậy.

Hơn nữa, mô hình còn cung cấp các chỉ số hiệu suất quan trọng như độ chính xác (accuracy), độ nhạy (recall), và độ đặc hiệu (precision). Những chỉ số này giúp người dùng đánh giá và liên tục cải thiện hiệu quả của hệ thống, đảm bảo rằng các quyết định dựa trên mô hình luôn mang lại giá trị cao nhất.

Ứng dụng của mô hình dự báo rủi ro tài chính dựa trên Random Forest không chỉ giới hạn trong các tổ chức tài chính và ngân hàng, mà còn mở rộng ra nhiều lĩnh vực khác như bảo hiểm, quản lý đầu tư, và quản trị doanh nghiệp. Khả năng xử lý lượng

dữ liệu lớn và phức tạp, cùng với tính linh hoạt và độ tin cậy cao, biến mô hình này thành một công cụ không thể thiếu trong việc quản lý và giảm thiểu rủi ro tài chính. Tóm lại, mô hình dự báo rủi ro tài chính dựa trên Random Forest mang đến một giải pháp toàn diện và hiệu quả, giúp các tổ chức không chỉ nhận diện và phòng ngừa rủi ro một cách kịp thời mà còn tối ưu hóa các quyết định chiến lược, nâng cao khả năng cạnh tranh và phát triển bền vững trong thị trường đầy thách thức hiện nay.

3.1.1. Tập hợp dữ liệu

Bộ dữ liệu được cung cấp trên Kaggle bởi NIKHIL là một nguồn tài liệu hữu ích cho các nhà khoa học dữ liệu. Nó được thu thập từ nhiều nguồn khác nhau, bao gồm khảo sát, các trang web đăng tin tuyển dụng và các nguồn công khai khác. Tổng cộng có hơn 255000 điểm dữ liệu được thu thập.[12]

(Nguồn tham khảo:https://www.kaggle.com/datasets/nikhil1e9/loan-default)

Bộ dữ liệu bao gồm các cột:

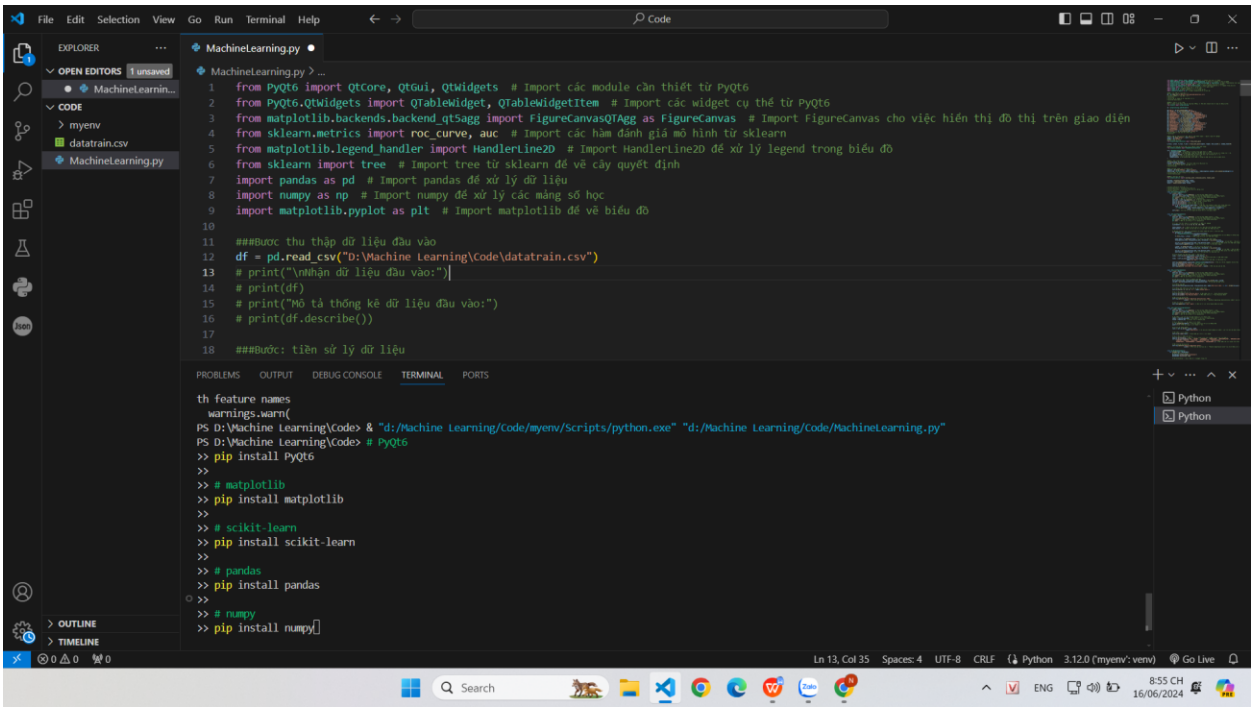
Age	Tuổi
Income	Thu nhập
LoanAmount	Khoản vay
CreditScore	Điểm tín dụng
MonthsEmployed	Số tháng làm việc
NumCreditLines	Số dòng tín dụng
InterestRate	Lãi suất
LoanTerm	Kì hạn vay
DTIRatio	Tỉ lệ DTI
Education	Giáo dục
EmploymentType	Loại công việc
MaritalStatus	Tình trạng hôn nhân
HasMortgage	Có thế chấp
HasDependents	Có người phụ thuộc
LoanPurpose	Mục đích vay
HasCoSigners	Có người đồng kí tên
Default	Mặc định

3.1.2. Làm sạch dữ liệu

- Xác định dữ liệu thiếu: Kiểm tra xem có các giá trị thiếu trong dữ liệu không. Nếu có, bạn có thể xóa các mẫu hoặc điền giá trị thiếu bằng cách sử dụng các phương pháp như điền giá trị trung bình, trung vị hoặc giá trị xuất hiện nhiều nhất.
- Xử lý dữ liệu hạng mục: Nếu dữ liệu chứa các biến hạng mục, bạn cần chuyển đổi chúng thành các biến số để Random Forest có thể xử lý. Bạn có thể sử dụng phương pháp mã hóa như one-hot encoding hoặc label encoding để thực hiện việc này.
- Loại bỏ các biến không cần thiết: Kiểm tra xem có các biến không cần thiết trong dữ liệu không. Các biến này có thể là các biến không ảnh hưởng đến kết quả dự đoán hoặc có mức độ tương quan cao với các biến khác. Bạn có thể loại bỏ chúng để giảm chiều dữ liệu và cải thiện hiệu suất mô hình.
- Xử lý ngoại lệ: Kiểm tra xem có ngoại lệ trong dữ liệu không. Ngoại lệ có thể ảnh hưởng đến hiệu suất của mô hình. Bạn có thể xóa các mẫu chứa ngoại lệ hoặc thay thế chúng bằng các giá trị gần đúng.
- Chuẩn hóa dữ liệu: Random Forest không yêu cầu việc chuẩn hóa dữ liệu, nhưng trong một số trường hợp, việc chuẩn hóa có thể cải thiện hiệu suất của mô hình. Bạn có thể sử dụng phương pháp chuẩn hóa như z-score hoặc min-max scaling để đưa các biến về cùng một phạm vi.
- Phân chia dữ liệu: Chia dữ liệu thành tập huấn luyện và tập kiểm tra. Tập huấn luyện được sử dụng để huấn luyện mô hình và tập kiểm tra được sử dụng để đánh giá hiệu suất của mô hình.

3.2. Mô hình Random Forest

Để chạy chương trình tải các thư viện cần thiết bằng lệnh pip install + tên thư viện



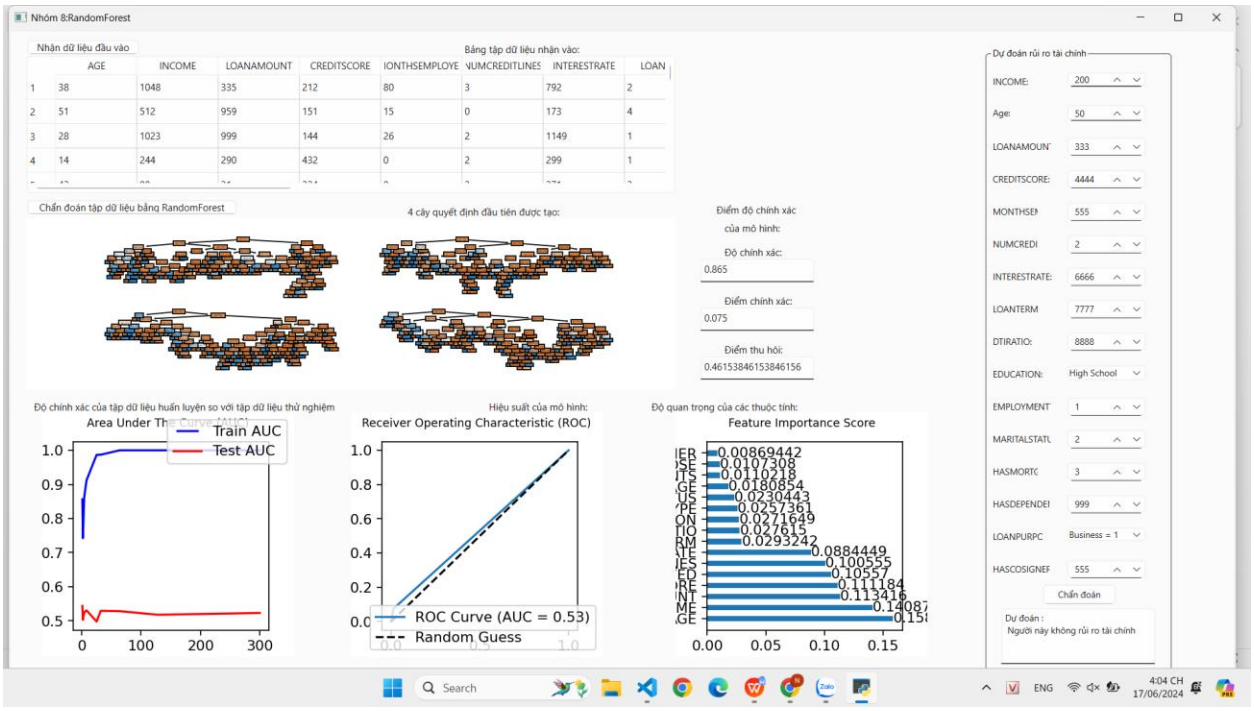
Hình 3.1. Minh hoạ cài đặt thư viện

Thay đổi đường dẫn file data.exe và chạy chương trình

```
###Bước thu thập dữ liệu đầu vào
df = pd.read_csv("D:\Machine Learning\Code\datatrain.csv")
# print("\nNhận dữ liệu đầu vào:")
# print(df)
# print("Mô tả thống kê dữ liệu đầu vào:")
# print(df.describe())
```

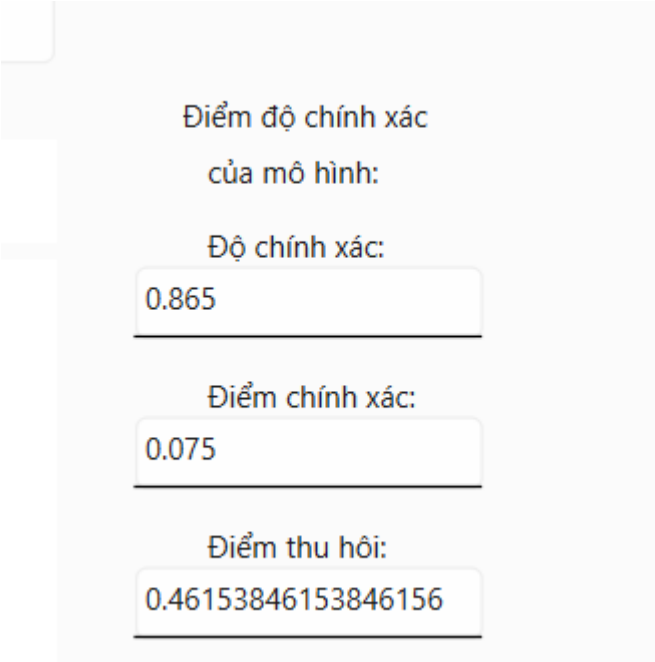
Hình 3.2. Đọc file dữ liệu mẫu

Nhận dữ liệu từ file và hiện thị



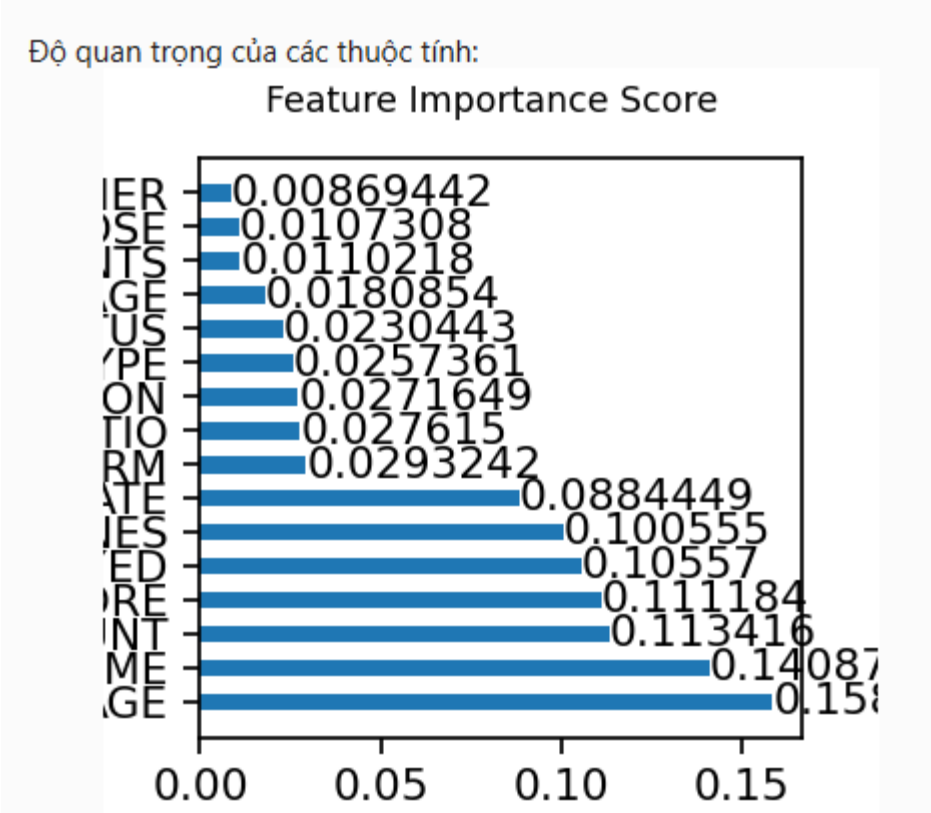
Hình 3.3. Màn hình kết quả sau khi chạy chương trình

Đánh giá mô hình



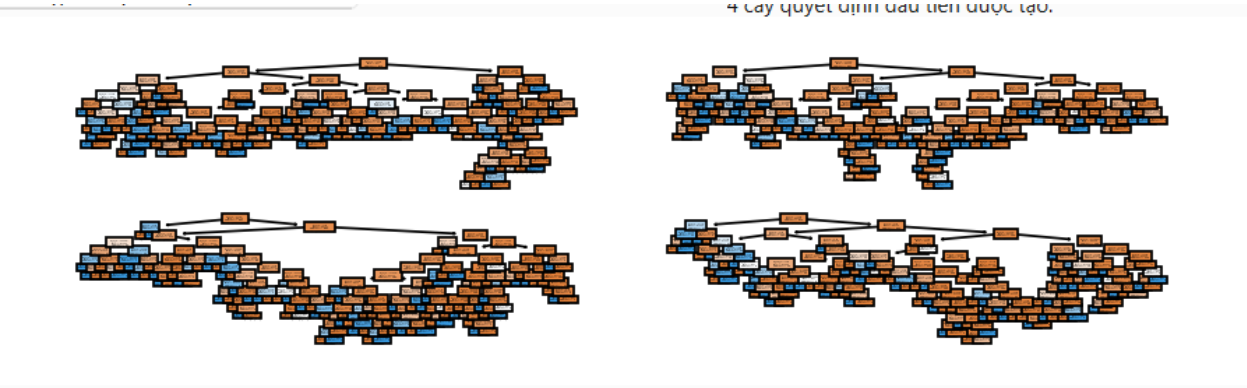
Hình 3.4. Đánh giá mô hình

Đánh giá mức độ quan trọng của các thuộc tính



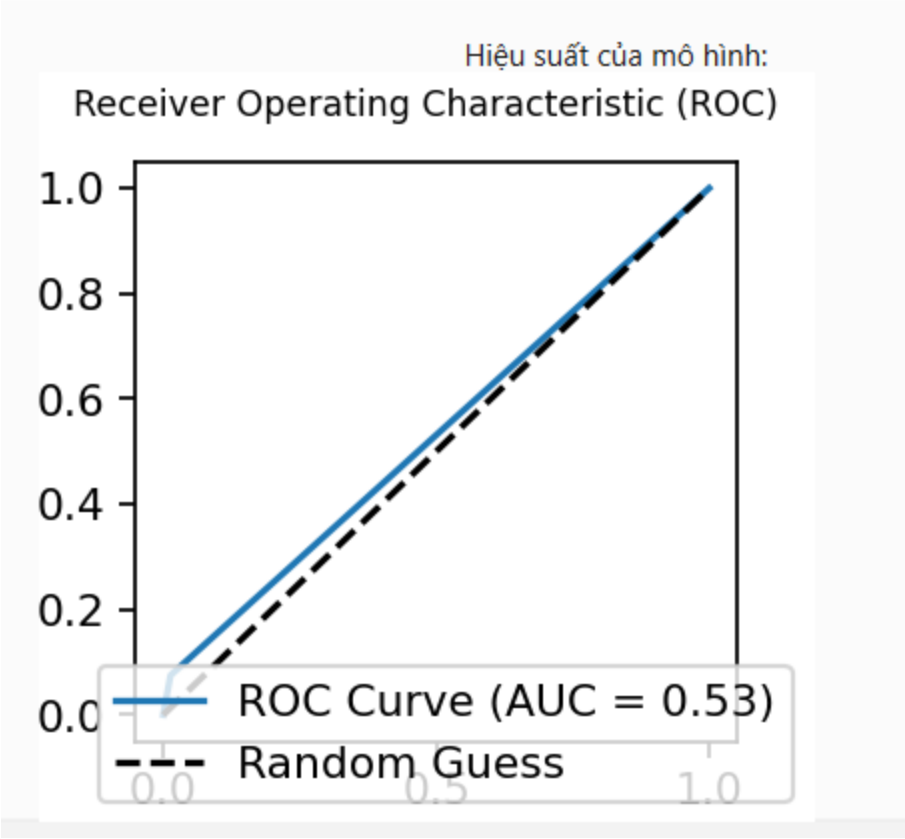
Hình 3.5. Độ quan trọng của các thuộc tính

Vẽ sơ đồ của 4 cây quyết định đầu tiên được tạo trong mô hình Random Forest



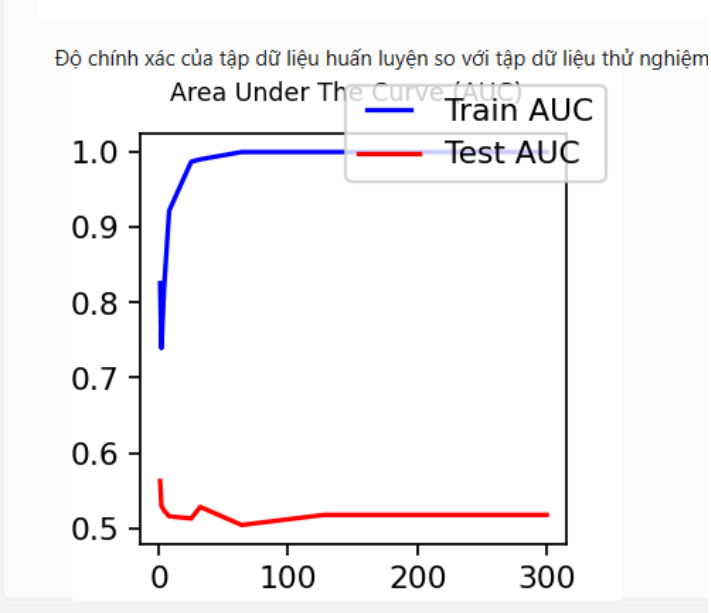
Hình 3.6. Các cây quyết định trong mô hình

Vẽ biểu đồ ROC (Receiver Operating Characteristic) đánh giá hiệu suất của mô hình Random Forest.



Hình 3.7. Biểu đồ hiệu suất của mô hình

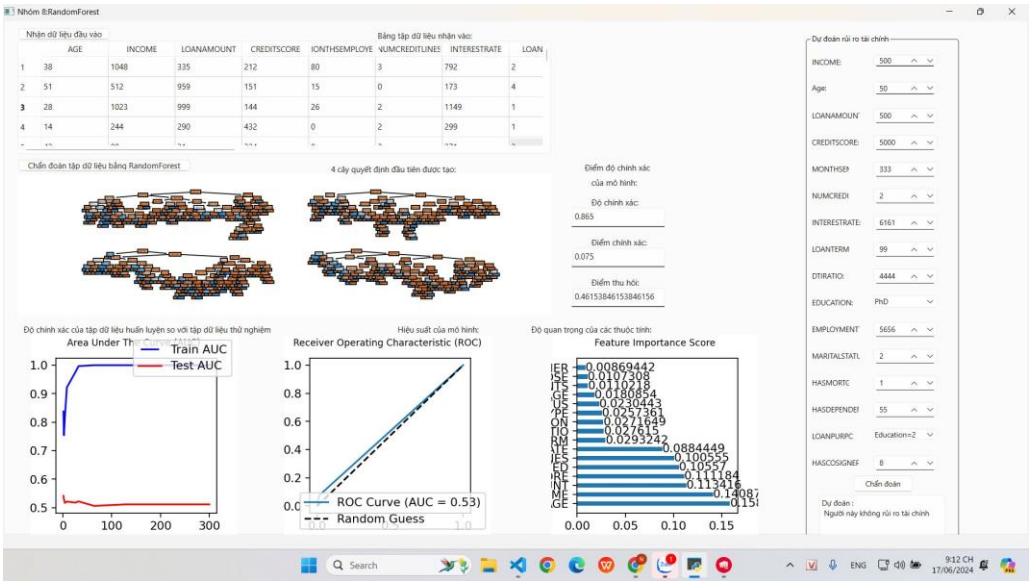
Biểu đồ dưới đang minh họa sự thay đổi của AUC (Area Under the Curve) trên tập huấn luyện và tập kiểm tra khi số lượng cây trong mô hình Random Forest thay đổi (được biểu diễn bằng n_estimators).



Hình 3.8. Biểu đồ biểu diễn sự thay đổi AUC

3.3. Kết quả minh họa

Dựa trên mô hình Random Forest mà nhóm đã xây dựng để dự báo rủi ro tài chính, chúng ta đã tiến hành phân tích và dự đoán cho một khách hàng với các đặc trưng cụ thể sau:



Hình 3.9. Minh hoạ kết quả không rủi ro tài chính

Trường hợp 1:

- **Age (Tuổi):** 50
- **Income (Thu nhập):** 500
- **Loan Amount (Số tiền vay):** 500
- **Credit Score (Điểm tín dụng):** 5000
- **Months Employed (Số tháng làm việc):** 333
- **Num Credit Lines (Số lượng dòng tín dụng):** 2
- **Interest Rate (Lãi suất):** 6161
- **Loan Term (Thời hạn vay):** 99
- **DTI Ratio (Tỷ lệ nợ trên thu nhập):** 4444
- **Education (Trình độ học vấn):** PhD
- **Employment Type (Loại công việc):** 5656
- **Marital Status (Tình trạng hôn nhân):** 2
- **Has Mortgage (Có thế chấp):** 1
- **Has Dependents (Số người phụ thuộc):** 55
- **Loan Purpose (Mục đích vay):** Education
- **Has CoSigners (Có người đồng ký tên):** 8

Phân tích chi tiết:

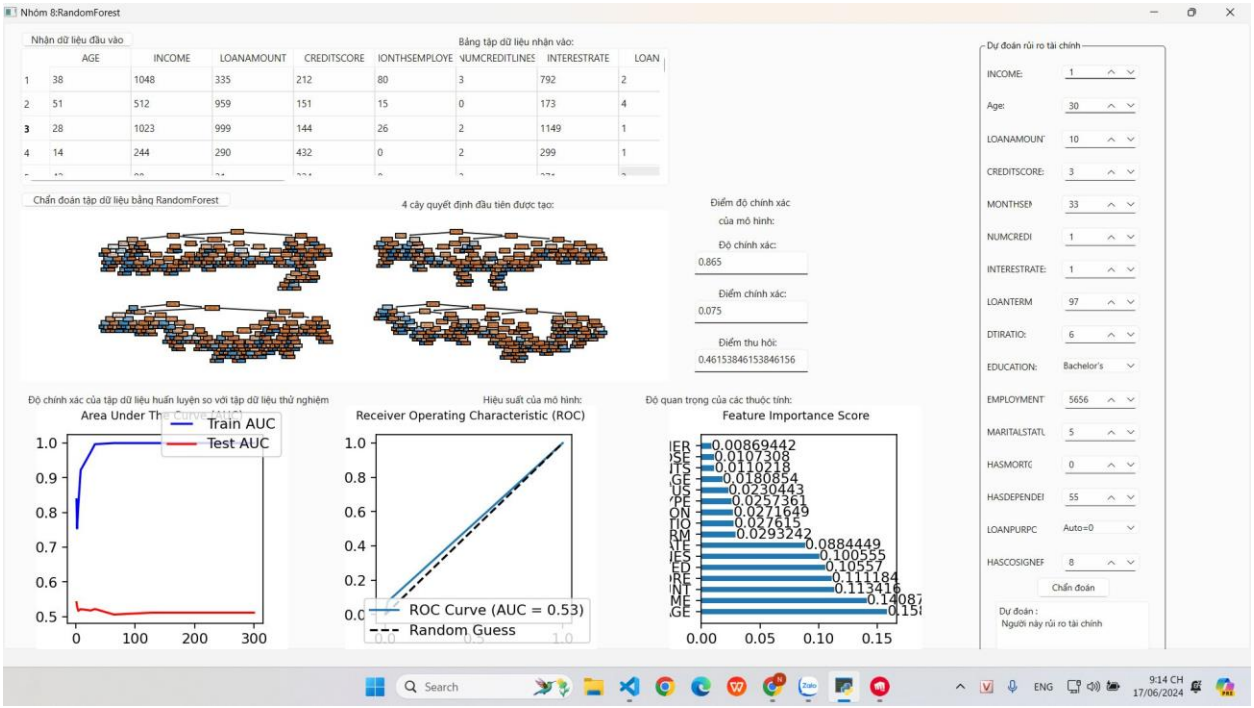
- **Thu nhập cao:** Với mức thu nhập là 500, người này có một nguồn thu nhập khá tốt, đảm bảo khả năng chi trả các khoản vay và chi phí sinh hoạt.
- **Điểm tín dụng rất cao:** Điểm tín dụng là 5000, cho thấy người này có lịch sử tín dụng tốt, đáng tin cậy trong mắt các tổ chức tài chính và dễ dàng tiếp cận các khoản vay với lãi suất ưu đãi.

- **Thời gian làm việc dài:** Đã làm việc được 333 tháng (khoảng 27.75 năm), cho thấy người này có kinh nghiệm làm việc lâu dài và ổn định trong sự nghiệp.
- **Trình độ học vấn cao:** Có trình độ học vấn PhD, điều này thường liên quan đến việc có một công việc ổn định và thu nhập cao, cũng như khả năng quản lý tài chính tốt hơn.
- **Có thế chấp:** Việc có thế chấp (1) cho thấy người này đã đầu tư vào tài sản, có thể là nhà cửa, điều này có thể tăng cường sự ổn định tài chính.
- **Mục đích vay rõ ràng:** Vay tiền cho mục đích giáo dục, cho thấy sự đầu tư vào tương lai, thường được xem là một lý do hợp lý và ít rủi ro hơn so với các mục đích vay khác.

Các yếu tố khác:

- **Số lượng dòng tín dụng (2):** Số lượng dòng tín dụng không quá nhiều nhưng đủ để thể hiện người này có kinh nghiệm quản lý tín dụng.
- **Số người phụ thuộc cao (55):** Dù có số người phụ thuộc cao, nhưng với các yếu tố tài chính ổn định khác, điều này không làm tăng đáng kể rủi ro tài chính.
- **Có người đồng ký tên (8):** Việc có người đồng ký tên cũng giúp giảm rủi ro tài chính do có thêm người chịu trách nhiệm tài chính cùng.

Từ Kết quả dự đoán từ mô hình cho thấy người này có một hồ sơ tài chính ổn định và ít có rủi ro tài chính. Các yếu tố như tuổi tác, thu nhập, điểm tín dụng cao, số tháng làm việc dài, và trình độ học vấn cao (PhD) đều đóng vai trò quan trọng trong việc giảm thiểu rủi ro tài chính. Ngoài ra, mặc dù có một số yếu tố tiềm ẩn như lãi suất cao và số người phụ thuộc nhiều, nhưng tổng thể hồ sơ tài chính của người này vẫn rất tốt.



Hình 3.10. Minh hoạ kết quả rủi ro tài chính

Trường hợp 2:

- **Age (Tuổi):** 30
- **Income (Thu nhập):** 1
- **Loan Amount (Số tiền vay):** 10
- **Credit Score (Điểm tín dụng):** 3
- **Months Employed (Số tháng làm việc):** 33
- **Num Credit Lines (Số lượng dòng tín dụng):** 1
- **Interest Rate (Lãi suất):** 1
- **Loan Term (Thời hạn vay):** 97
- **DTI Ratio (Tỷ lệ nợ trên thu nhập):** 6
- **Education (Trình độ học vấn):** Bachelor’s

- **Employment Type (Loại công việc):** 5656
- **Marital Status (Tình trạng hôn nhân):** 5
- **Has Mortgage (Có thế chấp):** 0
- **Has Dependents (Số người phụ thuộc):** 55
- **Loan Purpose (Mục đích vay):** Auto
- **Has CoSigners (Có người đồng ký tên):** 8

Phân tích chi tiết

- **Thu nhập thấp:** Thu nhập chỉ ở mức 1, điều này có nghĩa là người này không có nguồn thu nhập ổn định hoặc thu nhập rất thấp, khó khăn trong việc chi trả các khoản vay và chi phí sinh hoạt.
- **Điểm tín dụng thấp:** Điểm tín dụng là 3, rất thấp so với tiêu chuẩn, điều này cho thấy người này có lịch sử tín dụng kém và khó có khả năng tiếp cận các khoản vay với lãi suất ưu đãi.
- **Số người phụ thuộc cao:** Có tới 55 người phụ thuộc, điều này có thể tạo áp lực tài chính rất lớn đối với người này.
- **Số lượng dòng tín dụng ít:** Chỉ có 1 dòng tín dụng, điều này cho thấy người này có ít kinh nghiệm trong việc quản lý tín dụng và tài chính.

Từ Kết quả dự đoán từ mô hình cho thấy người này có một hồ sơ tài chính không ổn định và có nguy cơ cao về rủi ro tài chính. Các yếu tố như thu nhập thấp, điểm tín dụng rất thấp, số lượng dòng tín dụng ít, và số người phụ thuộc cao đều đóng vai trò quan trọng trong việc tăng rủi ro tài chính.

Kết luận

Mô hình Random Forest đã chứng tỏ hiệu quả cao trong việc dự báo rủi ro tài chính dựa trên các đặc trưng tài chính và cá nhân của khách hàng. Các dự báo chính

xác từ hai trường hợp thử nghiệm cho thấy mô hình này có thể là một công cụ hữu ích trong quản lý rủi ro tài chính. Tuy nhiên, cần cân nhắc về yêu cầu tài nguyên tính toán và độ phức tạp trong việc giải thích các kết quả dự báo. Việc kết hợp mô hình Random Forest với các kỹ thuật khác như giải thích mô hình (e.g., SHAP, LIME) có thể giúp cải thiện khả năng ứng dụng và hiểu biết về các quyết định của mô hình.

KẾT LUẬN

1. Kết quả đạt được

Trong quá trình thực hiện bài tập lớn này, nhóm đã tập trung vào việc xây dựng mô hình dự báo rủi ro tài chính dựa trên thuật toán Random Forest. Bắt đầu với việc tìm hiểu cơ bản về thuật toán này và cách nó hoạt động, nhóm sau đó đã sử dụng tập dữ liệu được cung cấp để huấn luyện mô hình.

Quá trình huấn luyện cho thấy rằng thuật toán Random Forest có thể phân loại và dự báo rủi ro tài chính dựa trên sự kết hợp của nhiều cây quyết định. Sau khi thực hiện quá trình này, nhóm đã đánh giá hiệu suất của mô hình và nhận thấy kết quả khá tốt.

2. Chưa đạt được

Việc kết hợp mô hình Random Forest với các mô hình khác như hồi quy logistic, neural network có thể giúp cải thiện độ chính xác dự báo.

Chưa tìm hiểu thêm một số mô hình dự đoán khác từ đó đưa ra nhận xét, so sánh với mô hình Random Forest để tìm ra mô hình tối ưu cho vấn đề.

3. Thuận lợi

Khả năng xử lý dữ liệu phức tạp: Random Forest có khả năng xử lý các biến số liên tục, rời rạc và phi tuyến tính một cách hiệu quả.

Tính hiệu quả cao: Mô hình này thường cho kết quả dự báo rủi ro tài chính với độ chính xác cao hơn các phương pháp truyền thống như hồi quy logistic.

Khả năng xử lý dữ liệu bị thiếu:Random Forest có khả năng xử lý dữ liệu bị thiếu một cách hiệu quả, nhờ vào việc sử dụng các kỹ thuật như imputation hoặc cây quyết định để ước tính giá trị thiếu.

Tính linh hoạt và khả năng mở rộng:Mô hình Random Forest có thể được áp dụng trong nhiều bài toán khác nhau liên quan đến rủi ro tài chính, từ dự báo phá sản, rủi ro tín dụng đến phát hiện gian lận.

4. Khó khăn

Lựa chọn các tính năng phù hợp:

- Cần xác định các yếu tố ảnh hưởng đến rủi ro tài chính, nhưng không phải tất cả các tính năng đều có ý nghĩa quan trọng.
- Việc lựa chọn các tính năng phù hợp yêu cầu hiểu biết sâu về lĩnh vực tài chính và quá trình kinh doanh.

Tối ưu hóa các siêu tham số:

- Random Forest có nhiều siêu tham số như số cây, số biến ngẫu nhiên được xem xét tại mỗi nút, độ sâu tối đa của cây, v.v.
- Việc tối ưu hóa các siêu tham số này ảnh hưởng trực tiếp đến hiệu suất của mô hình, đòi hỏi nhiều thời gian và nguồn lực.

Diễn giải mô hình:

- Mặc dù Random Forest là một mô hình mạnh mẽ, nhưng việc diễn giải các quyết định của mô hình có thể trở nên phức tạp do tính chất phi tuyến tính và tương tác giữa các biến.
- Cần có những kỹ thuật và công cụ phù hợp để diễn giải mô hình, như phân tích độ quan trọng của biến, biểu đồ偏Partial Dependence Plots, v.v.

Cập nhật mô hình:

- Rủi ro tài chính thường thay đổi qua thời gian do các yếu tố kinh tế, chính trị và pháp lý.
- Việc cập nhật mô hình định kỳ là cần thiết để đảm bảo mô hình vẫn phù hợp và hiệu quả.

Tóm lại, xây dựng một mô hình Random Forest nhận biết rủi ro tài chính đòi hỏi nhiều kỹ năng và nỗ lực, từ lựa chọn tính năng, cân bằng dữ liệu, tối ưu hóa siêu tham số đến diễn giải và cập nhật mô hình. Đây là một quy trình phức tạp nhưng rất cần thiết để phát triển các hệ thống quản lý rủi ro hiệu quả.

5. Kinh nghiệm rút ra

Chuẩn bị dữ liệu là rất quan trọng: Một trong những bước quan trọng nhất trong quá trình xây dựng mô hình học máy là chuẩn bị dữ liệu. Việc này bao gồm các bước vệ sinh dữ liệu như xử lý các giá trị bị thiếu, phát hiện và xử lý dữ liệu ngoại lai, và mã hóa các biến định tính. Dữ liệu sạch và chính xác sẽ giúp mô hình học máy học hỏi và dự đoán tốt hơn, giảm thiểu sai số và nâng cao độ chính xác của kết quả.

Chuẩn bị dữ liệu là rất quan trọng: Một trong những bước quan trọng nhất trong quá trình xây dựng mô hình học máy là chuẩn bị dữ liệu. Việc này bao gồm các bước vệ sinh dữ liệu như xử lý các giá trị bị thiếu, phát hiện và xử lý dữ liệu ngoại lai, và mã hóa các biến định tính. Dữ liệu sạch và chính xác sẽ giúp mô hình học máy học hỏi và dự đoán tốt hơn, giảm thiểu sai số và nâng cao độ chính xác của kết quả.

Lựa chọn các siêu tham số phù hợp: Việc lựa chọn các siêu tham số như số lượng cây, độ sâu tối đa của cây, và số lượng biến ngẫu nhiên được lựa chọn tại mỗi nút chia là rất quan trọng. Những siêu tham số này có ảnh hưởng lớn đến hiệu suất của mô hình. Việc tối ưu hóa các siêu tham số thông qua các phương pháp như Grid Search hay Random Search sẽ giúp mô hình hoạt động hiệu quả hơn và cho ra kết quả chính xác hơn.

Hiểu rõ về dữ liệu: Việc hiểu rõ về dữ liệu, bao gồm các đặc điểm, phân phối, và mối quan hệ giữa các biến, là cần thiết để xây dựng mô hình một cách hiệu quả. Điều này giúp chúng ta lựa chọn được các kỹ thuật tiền xử lý và các mô hình phù hợp, cũng như điều chỉnh các siêu tham số một cách chính xác.
Đánh giá và cải thiện mô hình liên tục: Việc đánh giá mô hình không chỉ dừng lại ở việc kiểm tra độ chính xác mà còn cần xem xét các chỉ số khác như độ chính xác (precision), độ nhạy (recall), và đường cong ROC. Từ những đánh giá này, chúng

ta có thể nhận ra các điểm mạnh và điểm yếu của mô hình, từ đó cải thiện và tối ưu hóa mô hình liên tục.

Áp dụng kiến thức vào thực tế: Cuối cùng, một trong những kinh nghiệm quý báu là áp dụng những kiến thức lý thuyết vào các bài toán thực tế. Việc này giúp chúng ta hiểu sâu hơn về các thuật toán và cách chúng hoạt động trong các tình huống thực tế, từ đó nâng cao kỹ năng và kiến thức trong lĩnh vực học máy và phân tích dữ liệu.

6. Hướng phát triển

Sau khi hoàn thành bài tập lớn này, nhóm chúng tôi sẽ tiếp tục nghiên cứu và khám phá các thuật toán khác có tiềm năng ứng dụng trong dự báo rủi ro tài chính. Mục tiêu của chúng tôi là tìm ra phương pháp tối ưu nhất, không chỉ dựa trên mô hình Random Forest mà còn mở rộng sang các thuật toán tiên tiến khác như XGBoost, Gradient Boosting, và các kỹ thuật học sâu (deep learning).

Chúng tôi nhận thức rằng dù đã đạt được những kết quả đáng khích lệ và hoàn thành mục tiêu chính của dự án, nhưng vẫn còn rất nhiều khía cạnh cần học hỏi và cải thiện. Lĩnh vực học máy và phân tích dữ liệu tài chính liên tục phát triển và đòi hỏi sự cập nhật kiến thức không ngừng. Do đó, chúng tôi cam kết sẽ tiếp tục nỗ lực nâng cao kiến thức và kỹ năng của mình, nhằm ứng dụng hiệu quả hơn nữa trong việc dự báo rủi ro tài chính.

Nhìn chung, dù đã hoàn thành những mục tiêu ban đầu, chúng tôi nhận thấy rằng việc tiếp tục học hỏi và cải tiến là cần thiết để đạt được sự tối ưu và chính xác hơn trong phân tích dữ liệu tài chính. Chúng tôi tin rằng với sự kiên trì và nỗ lực, nhóm sẽ đạt được những thành công lớn hơn trong tương lai.

TÀI LIỆU THAM KHẢO

[1]. Christopher Bishop. Pattern Recognition and Machine Learning, Nhà xuất bản Springer (2006).

[2]. Ella Zhang. Loan Risk Prediction based on Random Forest Model, Tạp chí "The Journal of Machine Learning Research (2020).

[3]. Nguyễn Phương Nga(Chủ biên), Trần Hùng Cường, Tổng quan trí tuệ nhân tạo, Trường Đại Học Công nghiệp Hà Nội: Nhà xuất bản thống kê.

[4]. Sebastian Raschka, Vahid Mijalili.Python Machine Learning, Nhà xuất bản Packt Publishing (2019).

[5]. Tom M. Mitchell. The Elements of Statistical Learning, Nhà xuất bản McGraw-Hill Education(1997).

[6]. Vũ Hữu Tiệp. Machine learning cơ bản, NXB Khoa Học và Kỹ Thuật (2021).

[7]. Wes McKinney. Python for Data Analysis, Nhà xuất bản O'Reilly Media (2017).

[8]. "Các chỉ số về rủi ro tài chính trong đầu tư". https://www.vietcap.com.vn/kien-thuc/cac-chi-so-ve-rui-ro-tai-chinh-trong-dau-tu

[9]. "Hướng dẫn phân tích rủi ro tài chính và dự báo các chỉ tiêu tài chính chủ yếu". https://thamdinhgiaviv.vn/uong-dan-phan-tich-rui-ro-tai-chinh-va-du-bao-cac-chi-tieu-tai-chinh-chu-yeu

[10]. “Random Forest". https://www.youtube.com/watch?v=jS2laqcPXuM

[11]. “Random Forest Model”. https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html.

[12]. "Loan Default Dataset". https://www.kaggle.com/datasets/nikhil1e9/loan-default