



PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN Ý ĐỊNH NGHỈ VIỆC CỦA NHÂN VIÊN: TIẾP CẬN TỪ MÔ HÌNH HỌC MÁY VÀ HÀM Ý QUẢN TRỊ CHO DOANH NGHIỆP VIỆT NAM

VŨ THỊ KIM LIÊN¹, NGUYỄN NGỌC KHÁNH², DƯƠNG THỊ HIỂN THANH³

Tóm tắt

Nghiên cứu áp dụng các phương pháp học máy tiên tiến: Support Vector Machine, Gradient Boosting, Logistic Regression, Random Forest và Deep Learning để phân tích các yếu tố ảnh hưởng đến quyết định nghỉ việc của nhân viên. Kết quả phân tích định lượng cho thấy, mô hình Gradient Boosting đạt hiệu suất cân bằng nhất với chỉ số F1-Score cho nhóm “Nghỉ việc” là 42%. Thông qua phương pháp Trí tuệ nhân tạo có thể diễn giải và biểu đồ SHAP, nghiên cứu xác định các yếu tố có tác động mạnh nhất đến ý định nghỉ việc gồm: Cấp độ quyền chọn cổ phiếu, Tình trạng làm thêm giờ, Số công ty đã làm việc và Sự hài lòng với môi trường.

Từ khóa: Dự báo nghỉ việc, hàm ý quản trị, học máy

ANALYZING FACTORS INFLUENCING EMPLOYEES' TURNOVER INTENTION: A MACHINE LEARNING APPROACH AND MANAGERIAL IMPLICATIONS FOR VIETNAMESE ENTERPRISES

Summary

This study applies advanced machine learning methods: Support Vector Machine, Gradient Boosting, Logistic Regression, Random Forest, and Deep Learning, to analyze the factors influencing employees' turnover decisions. Quantitative analysis results reveal that the Gradient Boosting model achieves the most balanced performance, with an F1-score of 42% for the “Resignation” group. Using interpretable artificial intelligence techniques and SHAP visualization, the study identifies the most influential factors affecting turnover intention, including stock option level, overtime status, number of companies previously worked for, and satisfaction with the work environment.

Keywords: Turnover prediction, managerial implications, machine learning

Ngày nhận bài: 23/9/2025; Ngày hoàn thành biên tập: 25/10/2025; Ngày duyệt đăng: 10/11/2025

GIỚI THIỆU

Tỷ lệ nghỉ việc cao đang là một thách thức lớn tại thị trường lao động Việt Nam, gây ra chi phí đáng kể và ảnh hưởng tiêu cực đến văn hóa doanh nghiệp, với xu hướng ngày càng gia tăng trong các ngành thâm dụng lao động và công nghệ cao (Gallup, 2024). Nhận thấy các phương pháp truyền thống như khảo sát và mô hình hồi quy tuyến tính thường không nắm bắt được các mối quan hệ phi tuyến tính phức tạp, nghiên cứu này áp dụng khoa học dữ liệu và học máy để xây dựng các mô hình dự báo với độ chính xác cao hơn (Fallucchi và cộng sự, 2020) organisational aspects and people management. In recent years, attention has increasingly been paid to human resources (HR. Cụ thể, nghiên cứu sẽ: (1) Xây dựng và so sánh hiệu suất của 5 mô hình học máy; (2) Áp dụng kỹ thuật lựa chọn đặc trưng và SMOTE (Synthetic Minority Over-sampling Technique) để xử

lý dữ liệu; (3) Sử dụng phương pháp SHAP (Shapley Additive exPlanations) để diễn giải mô hình hoạt động tốt nhất, từ đó xác định và lượng hóa mức độ ảnh hưởng của từng yếu tố, làm cơ sở đề xuất các hàm ý quản trị thực tiễn.

CƠ SỞ LÝ THUYẾT VÀ MÔ HÌNH NGHIÊN CỨU

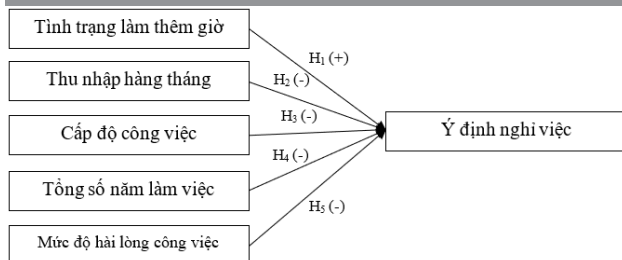
Cơ sở lý thuyết

Ý định nghỉ việc là một trong những chủ đề được nghiên cứu nhiều nhất trong lĩnh vực quản trị nhân sự và hành vi tổ chức. Lý thuyết kinh điển của (Mobley, 1977) đã chỉ ra rằng, sự không hài lòng trong công việc là tiền đề dẫn đến ý định và hành vi nghỉ việc. Các nghiên cứu truyền thống đã xác định nhiều yếu tố tác động thuộc về cá nhân, công việc và tổ chức, như sự cân bằng công việc-cuộc sống (Rani và cộng sự, 2023), căng thẳng (Lin và cộng sự, 2024), hay cơ hội phát triển (De Cássia và cộng sự, 2022). Tuy nhiên,

¹ ThS., ² PGS, TS., ³ TS. Trường Đại học Mở - Địa chất; Email: vuthikimlien@humg.edu.vn



HÌNH 1: MÔ HÌNH NGHIÊN CỨU ĐỀ XUẤT



Nguồn: Nhóm tác giả đề xuất

BẢNG 1: KẾT QUẢ MÔ HÌNH TỪ DỮ LIỆU GỐC

STT	Mô hình	Accuracy	Precision	Recall	F1-Score
1	Support Vector Machine (SVM)	0.90	0.87	0.27	0.41
2	Logistic Regression (LG)	0.90	0.68	0.40	0.50
3	Gradient Boosting (GB)	0.89	0.64	0.29	0.40
4	Random Forest (RF)	0.87	0.50	0.06	0.11
5	Deep Learning (DL)	0.86	0.45	0.35	0.40

Nguồn: Tính toán của nhóm tác giả

các phương pháp hồi quy tuyến tính truyền thống gặp khó khăn trong việc nắm bắt các mối quan hệ phi tuyến tính phức tạp. Do đó, xu hướng áp dụng học máy (ML) để dự báo nghỉ việc đã phát triển mạnh mẽ.

Nghiên cứu của Marín Díaz và cộng sự (2023) đi sâu vào ứng dụng phương pháp Trí tuệ nhân tạo có thể diễn giải (XAI) trong việc xác định khả năng luân chuyển nhân viên và đưa ra các giải pháp dựa trên dữ liệu để giải quyết vấn đề phức tạp này. Tương tự, nghiên cứu của Nguyễn Thị Thu Thảo (2024), cũng chỉ ra LR là mô hình tối ưu nhờ khả năng dự đoán chính xác và dễ diễn giải.

Hướng nghiên cứu thứ hai tập trung vào giải quyết các thách thức kỹ thuật, chủ yếu là vấn đề mất cân bằng dữ liệu (số người nghỉ việc luôn ít hơn số người ở lại). Kỹ thuật SMOTE được áp dụng rộng rãi để cải thiện khả năng nhận diện lớp thiểu số của mô hình (Anda và cộng sự, 2024). Nghiên cứu của Baydili và Tasci (2025) đã đề xuất một hệ thống kết hợp mô hình Transformer và SHAP để cung cấp các giải thích minh bạch, cho thấy các yếu tố như Mức độ hài lòng công việc và Số năm làm việc với quản lý hiện tại có ảnh hưởng lớn nhất. Tại Việt Nam, Luong Tien Vinh và Phan Thi Ngan (2025) cũng nhấn mạnh tính hiệu quả của các phương pháp dựa trên AI trong phân tích lực lượng lao động, cung cấp thông tin giá trị cho nhà quản lý.

Mô hình nghiên cứu

Dựa vào cơ sở lý thuyết nêu trên, nhóm tác giả xây dựng mô hình nghiên cứu như Hình 1.

Các giả thuyết nghiên cứu như sau:

H1: Tình trạng làm thêm giờ có tác động cùng chiều (+) đến Ý định nghỉ việc.

H2: Thu nhập hàng tháng có tác động ngược chiều (-) đến Ý định nghỉ việc.

H3: Cấp độ công việc có tác động ngược chiều (-) đến Ý định nghỉ việc

H4: Tổng số năm làm việc (kinh nghiệm) có tác động ngược chiều (-) đến Ý định nghỉ việc.

H5: Mức độ hài lòng với công việc có tác động ngược chiều (-) đến Ý định nghỉ việc.

Phương pháp nghiên cứu

Nghiên cứu này phân tích các yếu tố ảnh hưởng đến ý định nghỉ việc, tập trung vào 15 đặc trưng quan trọng nhất được lựa chọn từ bộ dữ liệu. Sử dụng phương pháp luận định lượng, áp dụng 5 mô hình học máy (SVM, GB, LR, RF và DL), kết hợp kỹ thuật SMOTE để xử lý mất cân bằng và XAI (SHAP) để diễn giải mô hình. Sử dụng nguồn dữ liệu thứ cấp của IBM trên nền tảng Kaggle với 1.470 quan sát.

Do dữ liệu mất cân bằng, việc đánh giá không chỉ dựa vào độ chính xác (Accuracy) mà tập trung vào các chỉ số có nguồn gốc từ ma trận nhầm lẫn: Recall, Precision và F1-Score để có cái nhìn toàn diện về hiệu suất thực tế của mô hình. Trong đó, 4 chỉ số cốt lõi của ma trận nhầm lẫn gồm: TP (True Positive): Số lượng nhân viên nghỉ việc được dự đoán đúng; TN (True Negative): Số lượng nhân viên không nghỉ việc được dự đoán đúng; FP (False Positive): Số lượng nhân viên nghỉ việc được dự đoán sai; FN (False Negative): Số lượng nhân viên không nghỉ việc được dự đoán sai.

Từ 4 chỉ số trên, xác định được các chỉ số đánh giá quan trọng:

- Accuracy (Độ chính xác): Độ chính xác càng cao càng tốt cho mô hình.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision (Độ chuẩn xác): Chỉ số này càng cao, số điểm mô hình dự đoán là Positive đều là Positive càng nhiều.

$$Precision = \frac{TP}{TP + FP}$$

- Recall: Recall càng cao, số điểm là positive bị bỏ sót càng ít.

$$Recall = \frac{TP}{TP + FN}$$

- F1-Score: là trung bình điều hòa của precision và recall.

$$F1_{score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

KẾT QUẢ NGHIÊN CỨU

Giai đoạn 1: Mô hình trên dữ liệu gốc

Phân tích mô hình trên dữ liệu gốc (Bảng 1) cho thấy, chỉ số Accuracy cao (86-90%) là một kết quả dễ gây hiểu lầm do sự mất cân bằng dữ liệu. Điểm yếu nghiêm trọng của các mô hình này là chỉ số Recall



rất thấp, đặc biệt là RF chỉ phát hiện được 6% số ca nghỉ việc thực tế, cho thấy sự thất bại trong việc xác định lớp thiểu số. Trong bối cảnh này, mô hình LR được xem là hiệu quả vì đạt được F1-Score cao nhất (0.50), thể hiện sự cân bằng tốt hơn giữa Precision (0.68) và Recall (0.40). Kết quả này khẳng định sự cần thiết phải áp dụng các kỹ thuật xử lý dữ liệu mất cân bằng như SMOTE để cải thiện hiệu suất dự báo.

Giai đoạn 2: Áp dụng lựa chọn đặc trưng

Kết quả cho thấy, việc lựa chọn đặc trưng đã xác định được các yếu tố quan trọng nhất ảnh hưởng đến quyết định nghỉ việc, trong đó Thu nhập hàng tháng, Tuổi và Tình trạng làm thêm giờ là các yếu tố có trọng số cao nhất. Sau khi chỉ giữ lại 15 đặc trưng quan trọng nhất từ RF, Accuracy của các mô hình giảm nhẹ. Đáng chú ý, chỉ số Recall của RF vẫn rất thấp (0.08), cho thấy việc lựa chọn đặc trưng đơn thuần không giải quyết được vấn đề mất cân bằng dữ liệu (Hình 2).

Như vậy, bước này cho thấy việc chỉ lựa chọn đặc trưng là không đủ. Mặc dù giúp mô hình đơn giản hơn, nhưng không giải quyết được vấn đề cốt lõi là sự mất cân bằng dữ liệu, vốn là nguyên nhân chính khiến các mô hình không học được các mẫu của nhân viên nghỉ việc.

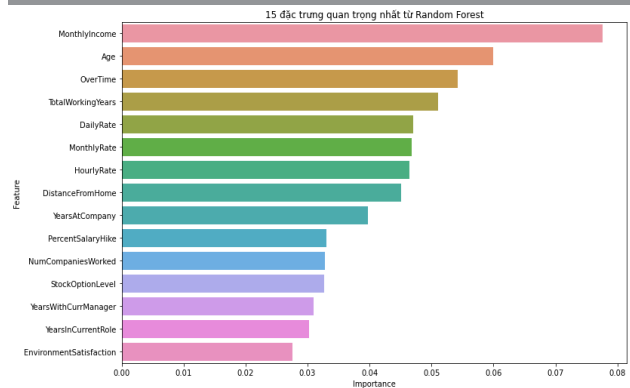
Giai đoạn 3: Áp dụng lựa chọn đặc trưng và SMOTE

Việc áp dụng kỹ thuật SMOTE đã giải quyết thành công vấn đề mất cân bằng dữ liệu, với kết quả nổi bật nhất là sự cải thiện vượt bậc về khả năng nhận diện các trường hợp nghỉ việc (chỉ số Recall). Cụ thể, số liệu Bảng 2 cho thấy, mô hình LR (Recall = 0.62) và SVM (Recall = 0.54) đã chứng tỏ hiệu quả cao trong việc xác định đúng hơn một nửa số nhân viên có ý định rời đi. Tuy nhiên, sự gia tăng này đi kèm với sự đánh đổi về chỉ số Precision, một hiện tượng phổ biến trong các bài toán phân loại. Xét về sự cân bằng tổng thể, mô hình GB được xác định là tối ưu nhất, thể hiện qua việc đạt được F1-Score cao nhất (0.42), là thước đo hài hòa giữa Precision và Recall. Kết quả này khẳng định SMOTE là một phương pháp hiệu quả để chuyển đổi mô hình từ độ chính xác lý thuyết sang giá trị ứng dụng thực tiễn.

Kết quả so sánh các mô hình bằng Ma trận nhầm lẫn

Kết quả (Bảng 3) cho thấy, GB được xác định là lựa chọn tối ưu và cân bằng nhất, đạt chỉ số F1-Score cao nhất (0.42), cho thấy sự hài hòa giữa khả năng phát hiện đúng các trường hợp nghỉ việc (Recall = 0.42) và độ chính xác của các dự đoán đó (Precision = 0.42). Trong khi đó, mô hình LR thể hiện khả năng “bắt” được nhiều trường hợp nhất với chỉ số Recall vượt trội (0.63), nhưng phải đánh đổi bằng Precision rất thấp (0.28), dẫn đến nhiều cảnh báo sai. Các mô hình RF và DL cho thấy hiệu quả kém nhất trong bài toán này.

HÌNH 2: BIỂU ĐỒ 15 ĐẶC TRƯNG TỪ RANDOM FOREST



Nguồn: Tính toán của nhóm tác giả

BẢNG 2: KẾT QUẢ MÔ HÌNH SAU KHI LỰA CHỌN ĐẶC TRƯNG VÀ SMOTE

STT	Mô hình	Accuracy	Precision	Recall	F1-Score
1	RD	0.86	0.44	0.23	0.30
2	GB	0.85	0.42	0.42	0.42
3	DL	0.82	0.32	0.33	0.33
4	SVM	0.79	0.33	0.54	0.41
5	LR	0.74	0.28	0.62	0.38

Nguồn: Tính toán của nhóm tác giả

BẢNG 3: SO SÁNH CHI TIẾT TỪ MA TRẬN NHẦM LẪN (SẮP XẾP THEO F1-SCORE VÀ RECALL)

Mô hình	Accuracy	TP (Phát hiện đúng)	FN (Bỏ sót)	Recall (lớp 1)	Precision (lớp 1)	F1-Score (lớp 1)
GB	0.85	20	28	0.42	0.42	0.42
SVM	0.79	26	22	0.54	0.33	0.41
LR	0.74	30	18	0.63	0.28	0.38
RF	0.86	11	37	0.23	0.44	0.30
DZL	0.80	15	33	0.31	0.27	0.29

Nguồn: Tính toán của nhóm tác giả

Diễn giải mô hình sử dụng SHAP

Trục ngang (Trục 0.0): Các điểm nằm bên phải trục 0.0 (giá trị SHAP dương) đều có tác động làm *tăng* khả năng dự đoán nghỉ việc. Các điểm nằm bên trái trục 0.0 (giá trị SHAP âm) đều có tác động làm giảm khả năng dự đoán nghỉ việc (tức là tăng khả năng ở lại).

Giá trị đặc trưng (Feature Value): Trên thang đo Feature value (bên phải biểu đồ): High (phía trên) tương ứng với giá trị đặc trưng cao (Thu nhập cao, tuổi cao, có làm thêm giờ); Low (phía dưới) tương ứng với giá trị đặc trưng thấp (Thu nhập thấp, tuổi thấp, không làm thêm giờ).

Phân tích từ SHAP đã làm sáng tỏ các yếu tố chính và cho phép đối chiếu lại với các giả thuyết ban đầu. Trái với các giả thuyết chỉ tập trung vào các yếu tố truyền thống, mô hình chỉ ra Cấp độ quyền chọn cổ phiếu là yếu tố quan trọng nhất, sau đó là Tình trạng làm thêm giờ và Số công ty đã làm việc.



Đối chiếu giả thuyết

H1 (Làm thêm giờ) và H2 (Thu nhập): Được xác nhận. Với H1 (OverTime): Hình 3 cho thấy các điểm dữ liệu ứng với giá trị đặc trưng cao (tức là Có làm thêm giờ) tập trung thành một cụm rõ rệt ở bên phải trục 0.0 (SHAP > 0). Điều này khẳng định việc làm thêm giờ đẩy dự đoán mạnh về phía nghỉ việc. Với H2 (MonthlyIncome): Tương tự, các điểm dữ liệu ứng với giá trị đặc trưng thấp (tức là Thu nhập thấp) có giá trị SHAP dương (nằm rải rác bên phải trục 0.0), làm tăng khả năng nghỉ việc. Ngược lại, các điểm ứng với Thu nhập cao (giá trị đặc trưng cao) lại tập trung bên trái (SHAP < 0), cho thấy xu hướng nghỉ việc cao hơn.

H4 (Tổng số năm làm việc): Được xác nhận. Nhân viên có Tổng số năm làm việc ít (ứng với giá trị đặc trưng thấp) có giá trị SHAP dương (tập trung ở bên phải trục 0.0), cho thấy nhóm này có xu hướng nghỉ việc cao hơn.

H3 (Cấp độ công việc) và H5 (Hài lòng công việc): Không được xác nhận là yếu tố hàng đầu. Các biến JobLevel (Cấp độ công việc) và JobSatisfaction (Hài lòng công việc) không xuất hiện trong top những yếu tố có ảnh hưởng mạnh nhất theo biểu đồ SHAP. Thay vào đó, biến EnvironmentSatisfaction (Sự hài lòng với môi trường) lại cho thấy tác động rõ rệt hơn (xếp thứ 5 về mức độ quan trọng).

Các phân tích trên cho thấy, tác động 2 chiều của Quyền chọn cổ phiếu:

Cấp độ quyền chọn cổ phiếu có tác động hai chiều bất ngờ: nhân viên với cấp độ thấp (mức 0 hoặc 1, tương ứng giá trị đặc trưng Low trên thang đo) có xu hướng ở lại (SHAP âm), trong khi những người có cấp độ cao lại có xu hướng nghỉ việc (SHAP dương). Điều này ngụ ý rằng khi quyền chọn đã mang lại lợi ích tài chính, nó không còn là yếu tố ràng buộc nhân viên tìm kiếm cơ hội mới.

Cấp độ quyền chọn cổ phiếu là yếu tố có ảnh hưởng mạnh nhất đến dự đoán, dựa trên mức độ ảnh hưởng trung bình (trị tuyệt đối của SHAP). Các yếu tố quan trọng tiếp theo là OverTime (Làm thêm

giờ) và NumCompaniesWorked (Số công ty đã làm việc), vượt qua các biến được giả thuyết ban đầu như JobLevel và JobSatisfaction.

KẾT LUẬN VÀ HÀM Ý QUẢN TRỊ

Nghiên cứu đã chứng minh thành công việc áp dụng học máy và XAI để xây dựng một mô hình dự báo nghỉ việc hiệu quả. Mô hình GB, sau khi xử lý dữ liệu bằng SMOTE, đã cho thấy hiệu suất đáng tin cậy. Quan trọng hơn, phân tích SHAP đã giải mã các yếu tố thực sự ảnh hưởng đến quyết định của nhân viên, bao gồm cả những yếu tố không được dự đoán trong giả thuyết ban đầu, đó là: Cấp độ quyền chọn cổ phiếu, Tình trạng làm thêm giờ, Số công ty đã làm việc và Sự hài lòng với môi trường. Từ những kết quả trên, nhóm tác giả đề xuất các hàm ý quản trị như sau:

Một là, tái cấu trúc quyền chọn cổ phiếu (ESOP): Doanh nghiệp cần thiết kế chính sách ESOP với lịch trình trao quyền kéo dài và gắn liền với các mục tiêu dài hạn của công ty. Điều này biến ESOP thành một công cụ giữ chân chiến lược, thay vì một phần thưởng tài chính ngắn hạn mà nhân viên có thể nhận rồi rời đi.

Hai là, các nhà quản lý cần chủ động rà soát lại khối lượng công việc, tối ưu hóa quy trình để tăng hiệu suất thay vì thời gian làm việc. Áp dụng chính sách làm việc linh hoạt, ghi nhận và khen thưởng dựa trên kết quả công việc, không phải số giờ có mặt tại văn phòng, sẽ giúp giảm tỷ lệ nghỉ việc đáng kể.

Ba là, xây dựng chính sách lương thưởng cạnh tranh và minh bạch: doanh nghiệp cần xây dựng hệ thống lương thưởng công bằng, minh bạch và có lộ trình tăng lương rõ ràng gắn với hiệu suất.

Bốn là, doanh nghiệp cần đầu tư vào việc xây dựng một văn hóa làm việc tích cực, nơi nhân viên được tôn trọng, lắng nghe và hỗ trợ. Các hoạt động gắn kết, các kênh phản hồi cởi mở và phong cách quản lý trao quyền sẽ là chìa khóa để nâng cao sự gắn bó của nhân viên. ■

Tài liệu tham khảo:

1. Anda, L., Jing, H., Yan, Y., and Qing, W. (2024). An Employee attrition prediction method base on smote and adaboost. *Computer Applications and Software*. <http://www.shcas.net/en/article/id/c3d25519-0d29-4e7c-a08d-5d39ebdd3053>.
2. Baydili, İ. T., and Tasci, B. (2025). Predicting Employee Attrition: XAI-Powered Models for Managerial Decision-Making. *Systems*, 13(7), 583.
3. Luong Tien Vinh và Phan Thi Ngan. (2025). Dự đoán tình trạng nghỉ việc của nhân viên bằng các phương pháp học máy. *Tạp chí Khoa học Lạc Hồng*, 1(20), 69-75. <https://doi.org/10.61591/jslhu.20.717>.
4. Marín Díaz, G., Galán Hernández, J. J., and Galdón Salvador, J. L. (2023). Analyzing Employee Attrition Using Explainable AI for Strategic HR Decision-Making. *Mathematics*, 11(22), 4677.
5. Mobley, W. H. (1977). Intermediate linkages in the relationship between job satisfaction and employee turnover. *Journal of Applied Psychology*, 62(2), 237-240. <https://doi.org/10.1037/0021-9010.62.2.237>.
6. Nguyễn Thị Thu Thảo. (2024). Nghiên cứu và dự đoán tỷ lệ nghỉ việc của người lao động bằng phương pháp học máy, luận án Tiến sĩ. <https://digital.lib.ueh.edu.vn/handle/UEH/74194>.