Dự đoán khả năng nghỉ việc của nhân viên bằng máy học

Nguyễn Lê Tấn Quang, Võ Đức Dương, Phạm Anh Kiệt, Phan Trọng Nhân
Trường Đại học Công nghệ Thông tin
Đại học Quốc gia Thành phố Hồ Chí Minh
HCMC, Việt Nam
{21522512, 21521992, 21522260, 21522407}@gm.uit.edu.vn

***Tóm tắt nội dung*—Tỉ lệ nhân viên nghỉ việc cao là tình trạng đáng lo ngại cho một doanh nghiệp bởi nó bào mòn năng lực tài chính và khả năng phát triển của doanh nghiệp đó. Doanh nghiệp cần chuẩn bị trước cho khả năng đó bằng cách tìm hiểu rằng những nhân viên nào sẽ nghỉ việc và vì sao họ nghỉ việc. Máy học là một hướng tiếp cận mới nổi do tính hiệu quả của nó trong bài toán phân loại. Trong bài báo cáo này, bộ dữ liệu IBM Human Resource Analytic Employee Attrition được khai thác bằng nhiều phương pháp khai phá dữ liệu và trích xuất đặc trưng khác nhau. Các kiến thức này được sử dụng để làm sạch và cải thiện chất lượng dữ liệu. Tiếp đó, bộ dữ liệu được dùng để chọn ra các đặc tính tối ưu cho các loại mô hình máy học khác nhau và huấn luyện trên các mô hình đó. Các mô hình được chọn trải dài trên năm nhóm khác nhau: Mô hình tuyến tính, mô hinh phân cụm, mô hình dạng cây, SVM và mạng nơ-ron nhân tạo. Thực nghiệm cho thấy rằng các mô hình hồi quy Logistic, CatBoost và AdaBoost là các mô hình cho kết quả tốt nhất với độ chính xác có trọng số dao động từ 0.5985-0.6364 và AUC ROC dao động từ 0.7244-0.7445. Đây là kết quả ở mức ổn và cho thấy rằng khả năng nghỉ việc của nhân viên có thể dự đoán được dựa theo đặc điểm liên quan đến họ, nhưng cần thêm sự cải tiến để áp dụng được việc dự đoán này ra thực tế.**

I. Giới thiệu

Hiện nay, tình hình nghỉ việc của nhân viên đang ở mức đáng chú ý. Theo thống kê của Vieclam24h.vn - website việc làm phổ biến hàng đầu Việt Nam, khoảng 75% người lao động có dự định nhảy việc trong 6 tháng đầu năm 2023 [1]. Theo một báo cáo khác của bộ Nội Vụ, tổng số công chức, viên chức thôi việc từ giữa năm 2022 đến giữa năm 2023 là 18.991 người, trong đó chủ yếu ở nhóm viên chức sự nghiệp giáo dục - đào tạo và sự nghiệp y tế [2]. Có hai nguyên nhân chính dẫn đến nghỉ việc, phân loại dựa trên quan hệ nhân viên đối với nhà tuyển dụng: Nghỉ việc tự nguyện và nghỉ việc không tự nguyện. Với nghỉ việc tự nguyện, nhân viên chủ động xin nghỉ việc, có thể do muốn nghỉ hưu sớm hay chuyển sang nơi khác có mức điều kiện tốt hơn. Với nghỉ việc không tự nguyện, nhân viên bị sa thải do không đáp ứng nhu cầu, hoặc do các chiến lược phát triển của doanh nghiệp. Tỉ lệ nghỉ việc cao ảnh hưởng xấu đến doanh nghiệp, vì khi một lượng lớn nhân viên nghỉ việc, doanh nghiệp sẽ phải tổ chức tuyển dụng [3]. Chi phí tổ chức phỏng vấn, tuyển dụng, đào tạo rất đắt đỏ, ảnh hưởng đến nguồn lực tài chính của doanh nghiệp. Đặc biệt, với nghỉ việc tự nguyện, những nhân viên nghỉ việc thường là những nhân viên giỏi, việc họ nghỉ việc sẽ khiến doanh nghiệp mất đi nhân lực trình độ cao. Ngoài ra, với những doanh nghiệp có dây chuyền sản xuất liên tục và nối tiếp, việc nhân viên nghỉ việc đột ngột có thể làm gián đoạn, trì trệ khả năng sản xuất của doanh nghiệp.

Do những nguy hiểm tiềm tàng của tỉ lệ nghỉ việc cao mang lại, việc phải có những sự chuẩn bị kĩ càng từ phía doanh nghiệp là hết sức cần thiết. Trong đó, doanh nghiệp cần trả lời được hai câu hỏi. Đầu tiên, những nhân viên có đặc điểm như thế nào thì sẽ nghỉ việc trong thời gian gần? Và, với những đặc điểm cho sẵn, liệu một nhân viên sẽ nghỉ việc trong thời gian gần? Với câu hỏi đầu tiên, đầu ra mong đợi của doanh nghiệp là các đặc điểm sẽ làm nhân viên nghỉ việc, chẳng hạn như lương, thưởng, sự thách thức, sự tự do, áp lực,... Biết được những đặc điểm này sẽ giúp công ty thay đổi hay cải thiện nơi làm việc, chế độ đãi ngộ,... từ đó làm giảm được tỉ lệ nghỉ việc, giúp công ty giảm chi phí tuyển dụng và giữ được nhân lực giỏi. Với câu hỏi thứ hai, đầu ra mong đợi của doanh nghiệp là khả năng một nhân viên sẽ nghỉ việc. Biết được khả năng nghỉ việc sẽ giúp doanh nghiệp biết những đặc điểm cần cải thiện, thay đổi, hay chuẩn bị trước được nhân lực thay thế. Việc này giúp dây chuyền làm việc của công ty không bị trì trệ và giúp bộ phận tuyển dụng nắm được tình hình, từ đó có kế hoạch tuyển dụng tốt hơn.

Để trả lời được những câu hỏi trên, người ta thường thu thập dữ liệu với số lượng lớn sau đó rút trích đặc trưng từ dữ liệu. Với lượng dữ liệu cực kì lớn, việc rút trích đặc trưng bằng tay tỏ ra không hiệu quả cả về mặt thời gian và độ chính xác. Thay vào đó, người ta sử dụng máy học - một công nghệ hỗ trợ phân tích và dự đoán trên dữ liệu. Chẳng hạn, các mô hình máy học có thể tìm ra những quy luật đằng sau những thông số có được từ dữ liệu. Mặt khác, các mô hình máy học có thể tự động dự đoán tỉ lệ nghỉ việc ở tốc độ rất nhanh, độ chính xác cao và không bị mệt mỏi hay thiên kiến.

Trong bài báo cáo này, một phân tích sơ bộ về dữ liệu sẽ được triển khai để xem xét các đặc điểm của nhóm nhân viên sẽ nghỉ việc và sẽ không nghỉ việc. Sau đó, các mô hình máy học cổ điển trong nhiều nhóm khác nhau (mô hình tuyến tính, mô hinh phân cụm, mô hình dạng cây, SVM và mạng nơ-ron nhân tạo) sẽ được sử dụng để cố gắng giải quyết bài toán dự đoán khả năng nghỉ việc của nhân viên. Bộ dữ liệu được dùng là bộ IBM Human Resource Analytic Employee Attrition có từ Kaggle Dataset Repository.

Cấu trúc của các phần tiếp theo của bài báo cáo như sau: Phần II nói về bộ dữ liệu được dùng, phân tích sơ bộ tập dữ liệu và thực hiện tiền xử lý dữ liệu; Tiếp đó, Phần III nói về các mô

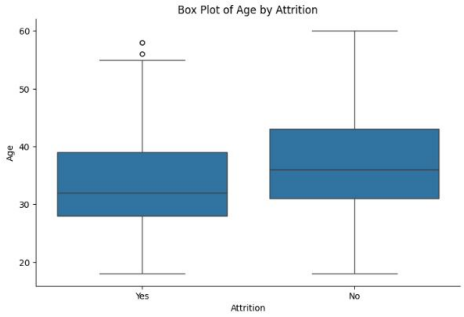
hình máy học được sử dụng, cài đặt thực nghiệm và độ đo; Sau đó, Phần IV nói về các thử nghiệm tinh chỉnh mô hình; Tiếp theo, Phần V phân tích các lỗi của mô hình và vạch ra hướng phát triển trong tương lai; Và cuối cùng, Phần VI là phần kết luận của bài báo cáo.

II. Dữ liệu

A. *Về bộ dữ liệu được dùng*

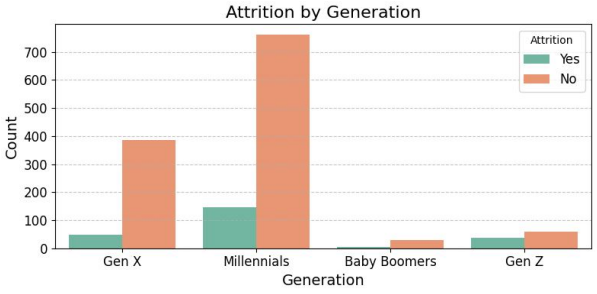
Bộ dữ liệu IBM Human Resource Analytic Employee Attrition (IBM-HRAEA) là bộ dữ liệu giả tưởng được tạo ra bởi các nhà khoa học ở IBM và được công khai trên Kaggle[1]. Bộ dữ liệu bao gồm 1470 mẫu với 35 thuộc tính, bao gồm các thông tin nhân khẩu học, lương, kinh nghiệm và mức độ hài lòng của nhân viên và thuộc tính mục tiêu chỉ ra nhân viên có nghỉ việc không. Trong 1470 nhân viên, có 86% nhân viên không nghỉ việc và 14% nhân viên nghỉ việc, chỉ ra rằng dữ liệu có sự mất cân bằng rất lớn.

B. *Ảnh hưởng của một số thuộc tính lên khả năng nghỉ việc*



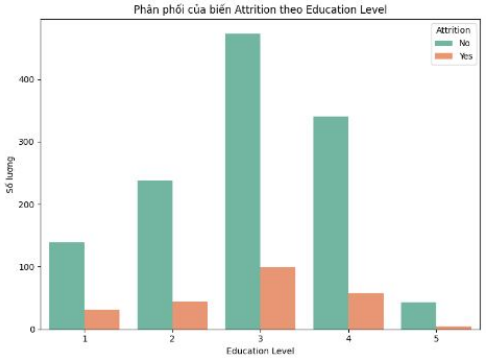
Hình 1. Biểu đồ Box tuổi nghỉ việc

1) Tuổi và thế hệ: Về tuổi, những người có tuổi cao hơn có xu hướng ít ra đi khỏi tổ chức hơn những người trẻ tuổi.
Thế hệ được xác định bằng tuổi, trong đó có 4 thế hệ được xem xét: Baby Boomers (>56), Gen X (>40, <=56), Gen Y (>24, <=40), Gen Z (<=24).



Hình 2. Biểu đồ cột số lượng nghỉ việc theo thế hệ

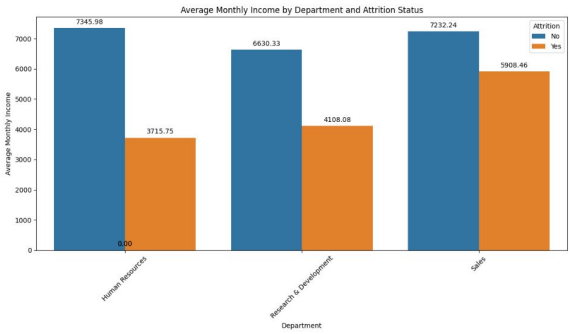
Từ biểu đồ trên, thấy được: Thế hệ Millennials chiếm số lượng lớn nhất trong tổng số nhân viên, cũng như số lượng nhân viên nghỉ việc; Gen X và Baby Boomers có số lượng nhân viên nghỉ việc ít hơn so với Millennials, mặc dù số lượng tổng cộng của Gen X cũng khá cao; Gen Z, mặc dù có số lượng nhân viên ít nhất, nhưng có tỷ lệ nghỉ việc tương đối cao so với tổng số lượng. Cụ thể, Gen Z có tỉ lệ nghỉ việc gần 0.4, trong khi các thế hệ khác có tỉ lệ nghỉ việc chỉ từ 0.1-0.16.



Hình 3. Biểu đồ cột số lượng nghỉ việc theo trình độ học vấn

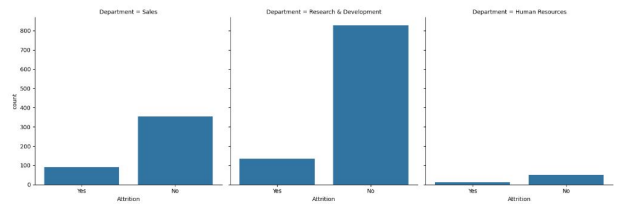
2) Trình độ học vấn: Biểu đồ trên cho thấy tỷ lệ nghỉ việc dường như không bị ảnh hưởng rõ ràng bởi trình độ học vấn vì tỷ lệ nghỉ việc luôn nhỏ hơn tỷ lệ ở lại ở tất cả các cấp học.

3) Thu nhập và phòng ban:



Hình 4. Biểu đồ cột lương trong theo thuộc tính nghỉ việc và phòng ban

Biểu đồ thể hiện rằng những nhân viên đã rời tổ chức có mức thu nhập trung bình thấp hơn so với những nhân viên vẫn ở lại. Điều này có thể cho thấy thu nhập có thể là một yếu tố quan trọng ảnh hưởng đến quyết định nghỉ việc của nhân viên.

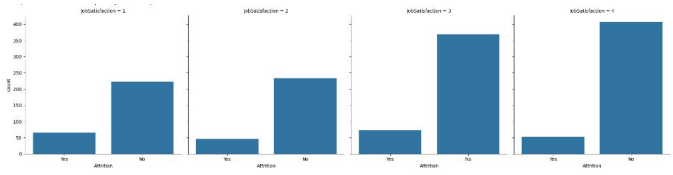


Hình 5. Biểu đồ cột số lượng nghỉ việc theo phòng ban

Biểu đồ cho biết rằng tỉ lệ nghỉ việc ở phòng nghiên cứu và phát triển thấp hơn so với hai phòng ban nhân lực và kinh doanh.

[1]https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

4) Giới: Tỉ lệ nghỉ việc của nam cao hơn nữ, tuy nhiên chênh lệch không nhiều (0.170 so với 0.148).



Hình 6. Biểu đồ cột số lượng nghỉ việc theo từng mức độ hài lòng về công việc

5) Mức độ hài lòng về công việc: Nhận xét rằng, ở các mức độ hài lòng về công việc cao hơn (người lao động hài lòng hơn với công việc của mình), tỷ lệ phần trăm của những người này nói muốn ra đi là thấp hơn.

6) Mức độ hài lòng về môi trường làm việc: Nhận xét rằng, với những người có cảm nhận tích cực hơn về môi trường làm việc có xu hướng ổn định hơn và ít có khả năng muốn rời đi khỏi tổ chức.

7) Mức độ tham gia vào công việc: Nhận xét rằng, những người tham gia vào công việc nhiều hơn có xu hướng ổn định và ít có khả năng muốn rời đi khỏi tổ chức.

8) Mức độ cân bằng giữa công việc - cuộc sống: Nhận xét rằng, những người có mức độ cân bằng giữa công việc và cuộc sống cao hơn có xu hướng ổn định hơn và ít có khả năng muốn rời đi khỏi tổ chức.

9) Mức độ hài lòng về mối quan hệ: Nhận xét rằng, những người có mức độ hài lòng về mối quan hệ cao hơn có xu hướng ổn định hơn và ít có khả năng muốn rời đi khỏi tổ chức.

C. Tiền xử lý dữ liệu

1) Chia dữ liệu: Tập kiểm định được chọn ngẫu nhiên 0.15 số mẫu từ bộ dữ liệu, giữ nguyên phân phối của bộ dữ liệu. Tập kiểm tra được chọn ngẫu nhiên 0.15 số mẫu từ bộ dữ liệu theo cách tương tự. Việc chia dữ liệu được thực hiện nhờ hàm train_test_split của thư viện sklearn.

2) Làm sạch dữ liệu: Làm sạch dữ liệu giúp loại bỏ những điểm hay thuộc tính của dữ liệu có khả năng cao ảnh hưởng xấu đến việc phân tích hay dữ đoán trên dữ liệu. Việc tồn tại các điểm hay thuộc tính "không sạch"có thể do sự bất cẩn hoặc thiếu khả năng chuyên môn trong nhập liệu hay thu thập thông tin, do tự nhiên vốn tồn tại những cá thể rất khác với những cá thể khác, do thiếu tương thích giữa người thu thập dữ liệu và máy tính,... Làm sạch dữ liệu bao gồm làm sạch điểm dữ liệu và làm sạch thuộc tính.

Về việc làm sạch điểm dữ liệu, dữ liệu không có những điểm trùng, thiếu hay ngoại lai nên không cần xử lý thêm (Riêng với xử lý ngoại lai, phương pháp IQR đánh giá rằng xấp xỉ 50% lượng dữ liệu là ngoại lai, lượng ngoại lai này quá lớn để bị loại bỏ nên dữ liệu được giữ nguyên).

Về việc làm sạch thuộc tính, dữ liệu có các thuộc tính EmployeeCount, Over18, StandardHours có phương sai quá thấp (ngưỡng là 0.1) nên bị loại bỏ. Thuộc tính EmployeeNumber cũng không quan trọng vì đây chỉ là đánh số các nhân viên nên cũng có thể loại bỏ. Sau đó, các giá trị phân loại được chuyển về kiểu số bằng Dummy Encoding. Mọi dữ liệu được chuẩn hoá về khoảng 0-1. Tiếp theo, dữ liệu được xử lý để loại bỏ sự đa cộng tuyến giữa các thuộc tính. Việc này giúp dữ liệu trở nên đơn giản, giúp cho mô hình học nhanh và tốt hơn. Đầu tiên, xét quan hệ giữa từng cặp thuộc tính, loại bỏ các thuộc tính có giá trị tuyệt đối tương quan lớn hơn 0.7 với một thuộc tính còn lại khác. Kết quả của quá trình này là, chín thuộc tính được loại bỏ. Tiếp theo, hệ số lạm phát phương sai (Variance inflation factor – VIF) được sử dụng để giải quyết vấn đề đa cộng tuyến. Hệ số này định lượng mức độ nghiêm trọng của đa cộng tuyến trong phân tích hồi quy bình phương nhỏ nhất bình thường. Một thuộc tính có VIF cao cho thấy rằng thuộc tính này có thể được suy ra từ các thuộc tính có sẵn khác. Ở bài báo cáo này, các thuộc tính có VIF > 10 được loại bỏ. Có bảy thuộc tính bị loại bỏ sau quá trình này: JobLevel, Age, Department_Research & Development, PercentSalaryHike, MonthlyIncome, HourlyRate, YearsAtCompany.

3) Xử lý mất cân bằng dữ liệu: Mất cân bằng dữ liệu có thể làm cho các mô hình gặp khó khăn trong việc học và có thể làm mô hình bị quá khớp. Về mặt đánh giá, mất cân bằng dữ liệu có thể gây ngộ nhận chất lượng mô hình. Chẳng hạn, trong phân loại nhị phân, độ chính xác cao có thể không đồng nghĩa với mô hình tốt, bởi có khả năng mô hình dự đoán thiên về lớp đa số, làm cho độ chính xác cao hơn mặc dù chất lượng thực sự của mô hình không tăng. Có nhiều hướng tiếp cận để xử lý mất cân bằng dữ liệu, nhưng bài báo cáo này sử dụng SMOTE.SMOTE hoạt động bằng cách chọn hai điểm dữ liệu cùng lớp, vẽ một đường nối hai điểm này và chọn một điểm dữ liệu ngẫu nhiên trên đường nối đó. Cụ thể, SMOTE sẽ bắt đầu với một điểm dữ liệu trong lớp thiểu số, chọn k điểm cùng lớp lân cận, chọn ngẫu nhiên một trong k điểm đó và thực hiện chọn điểm mới dựa trên điểm dữ liệu gốc và điểm được chọn ngẫu nhiên. Trong bài báo cáo này, k=10 được lựa chọn.

III. Phương pháp máy học

Nhiều phương pháp máy học được thử nghiệm trong bài báo cáo, nhằm khai thác và khám phá triệt để hiệu quả của các phương pháp này lên bộ dữ liệu đã trình bày. Trong đó, các phương pháp máy học mà bài báo cáo đề xuất có thể được chia thành năm nhóm chính: Mô hình tuyến tính, mô hinh phân cụm, mô hình dạng cây, SVM và mạng nơ-ron nhân tạo.

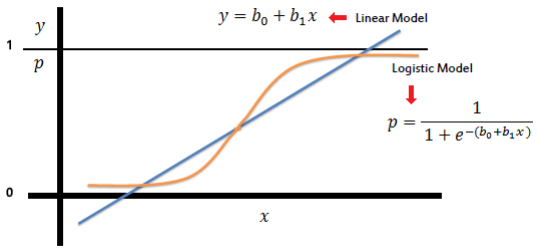
A. Các mô hình máy học

1) Mô hình tuyến tính: Trong nhóm các mô hình tuyến tính, mô hình hồi quy logistic được lấy làm đại diện. Mô hình hồi quy logistic giả định đầu ra của mô hình là một xác suất, và quan hệ giữa đầu vào và hàm mục tiêu có thể được xem như hàm Logistic [4]. Hàm logistic có dạng như sau:

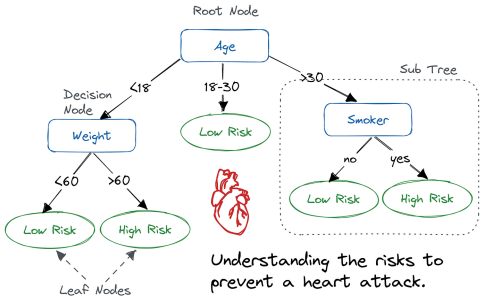
$$L = ln\frac{P(Y)}{1-P(Y)} = B_0 + B_1X_1 \quad (1)$$

Phân phối Logistic có dạng chữ S, có giá trị nằm trong khoảng từ 0 đến 1 khi B_1X_1 thay đổi từ âm vô cùng đến dương vô cùng. Việc ước lượng các tham số của công thức được dựa trên một phương pháp gọi là Maximum Likelihood. Mục tiêu

[2]https://www.saedsayad.com/logistic$_r egression.htm$



Hình 7. Ví dụ về hồi quy logistic [2]



Hình 8. Ví dụ về cây quyết định [3]

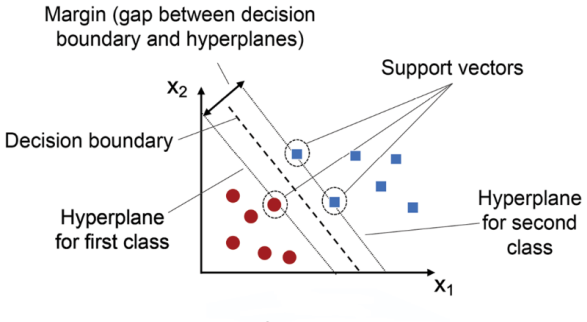
cuối cùng là tối thiểu hoá cross-entropy giữa phân phối xác suất dự đoán và phân phối lớp sử dụng một phương pháp tối ưu hoá, chẳng hạn như gradient descent.

2) Mô hình phân cụm: Trong nhóm các mô hình tuyến tính, mô hình K-Nearest Neighbors (K-NN) được lấy làm đại diện. K-NN là một mô hình học "lười biếng"(lazy learning), bởi vì nó chỉ ghi nhớ tập huấn luyện thay vì xử lí và rút trích đặc trưng như các mô hình khác. Mô hình này giả định rằng các điểm dữ liệu có quan hệ không gian với nhau, và những điểm dữ liệu gần nhau sẽ có đặc điểm tương tự nhau. K-NN yêu cầu đầu vào K được xác định bởi người dùng, rồi dự đoán bằng cách lấy dự đoán của K điểm dữ liệu gần nhất trong tập huấn luyện rồi tổng hợp kết quả. Ngoài giá trị K, K-NN còn yêu cầu một thang đo khoảng cách giữa hai điểm, mà thường là khoảng cách Euclid, khoảng cách Manhattan hay khoảng cách Minkowski.

3) Mô hình dạng cây: Trong nhóm các mô hình dạng cây, các mô hình được thử nghiệm bao gồm cây quyết định, Random Forest, CatBoost, LGBM, AdaBoost, XGBoost. Cây quyết định là một đồ thị dạng cây thể hiện việc đưa ra lựa các lựa chọn và kết quả các lựa chọn đó. Mỗi nút trong của cây quyết định thể hiện một câu hỏi hoặc một chỉ dẫn để đưa ra quyết định, mỗi cạnh trong cây thể hiện một sự lựa chọn và các nút lá thể hiện kết quả của lựa chọn. Cây quyết định giả định rằng toàn bộ tập huấn luyện nằm ở nút gốc, sau đó tập huấn luyện được chia thành các phần nhỏ hơn ở các nút con bằng đệ quy cho đến nút lá. Các tiêu chuẩn chia là các lựa chọn (là các ngưỡng với dữ liệu liên tục). Các tiêu chuẩn chia và thứ tự chia được quyết định bằng các thuật toán như ID3, C4.5, CART, CHAID,... Các loại mô hình khác sử dụng nhiều cây quyết định rồi tổng hợp kết quả, khắc phục được sự không ổn định và dễ bị quá khớp của cây quyết định.

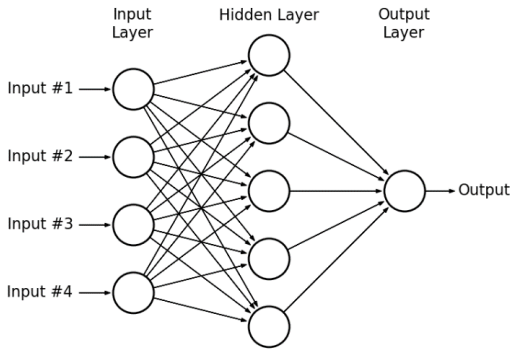
4) SVM: SVM là mô hình phi xác suất được dùng cho các bài toán phân loại và hồi quy. SVM huấn luyện bằng cách phân các điểm dữ liệu bằng các biên giới là các siêu phẳng theo nguyên tắc, các điểm nên được phân đúng lớp và càng xa biên càng tốt. SVM còn sử dụng các kernel function cho các bài toán phi tuyến tính hoặc/và khó xác định biên giới.

5) Mô hình nơ-ron nhân tạo: Mô hình nơ-ron nhân tạo là một nhóm các thuật toán được lấy cảm hứng từ chính các vận hành của não người, trong đó cấu trúc mô hình bao gồm các



Hình 9. Ví dụ về SVM [4]

nơ-ron, các lớp nơ-ron và các liên kết. Trong nhóm các mô hình nơ-ron nhân tạo, mô hình Multilayer Perceptron (MLP) được lấy làm đại diện bởi tính đơn giản của nó. MLP là một mạng nơ-ron đơn giản chỉ gồm các lớp (bao gồm các nút) và các hàm kích hoạt, trong đó các nút giữa hai lớp liên tiếp được nối đầy đủ, một chiều với nhau từ đầu vào đến đầu ra. MLP học bằng các thuật toán lan truyền tiến và lan truyền ngược.



Hình 10. Ví dụ về Multilayer Perceptron[5]

B. Các siêu tham số

1) Mô hình tuyến tính: Mô hình hồi quy logistic có ba siêu tham số đáng chú ý:

[3]https://www.datacamp.com/tutorial/decision-tree-classification-python
[4]https://vitalflux.com/classification-model-svm-classifier-python-example/

[5]https://www.researchgate.net/figure/A-hypothetical-example-of-Multilayer-Perceptron-Network_fig4_303875065

- C: hệ số chính quy hoá, C càng cao thì trọng số dữ liệu huấn luyện càng cao, ngược lại thì trọng số yếu tố chính quy hoá càng cao. C trong bài báo cáo được thử nghiệm với ba giá trị [0.1, 1, 10].
- penalty: Loại yếu tố chính quy hoá được sử dụng. Các loại chính quy hoá được thử nghiệm là [None (không sử dụng chính quy hoá), 'l1', 'l2']
- solver: Thuật toán được sử dụng để tối ưu hoá mô hình. Các thuật toán được thử nghiệm là ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']

2) Mô hình phân cụm: Mô hình phân cụm có các siêu tham số đáng chú ý:

- n_neighbors: Số lân cận (số k). Các giá trị được thử nghiệm là [3, 5, 10].
- weights: Trọng số của các lân cận. Các giá trị được thử nghiệm là ['uniform', 'distance'].
- metric: Độ đo khoảng cách. Các loại độ đo được thử nghiệm là ['euclidean', 'manhattan'].

3) Mô hình dạng cây: Mô hình dạng cây có các siêu tham số đáng chú ý:

- criterion: Độ đo chất lượng của một điểm chia. Các thuật toán được thử nghiệm là ['gini', 'entropy'].
- max_depth, min_samples_split, min_samples_leaf: Các siêu tham số để chỉ định hình dáng của cây, lần lượt là độ sâu tối đa, số mẫu tối thiểu để phân chia và số mẫu tối thiểu ở nút lá. Thử nghiệm với nhiều giá trị khác nhau xoay quanh giá trị mặc định.
- max_features: Lượng đặc trưng tối đa lúc xét tìm điểm chia. Các giá trị được thử nghiệm là [None (không giới hạn lượng đặc trưng), 'sqrt', 'log2'].

Không phải mô hình nào cũng có các siêu tham số trên, đặc biệt là các mô hình kết hợp. Mô hình nào không có siêu tham số nào thì bỏ qua siêu tham số đó.

4) SVM: SVM có các siêu tham số đáng chú ý:

- C: hệ số chính quy hoá (như ở mô hình tuyến tính). C trong bài báo cáo được thử nghiệm với ba giá trị [0.1, 1, 10].
- kernel: Loại kernel được sử dụng. Các loại được thử nghiệm là ['linear', 'poly', 'rbf', 'sigmoid'].

5) Mô hình nơ-ron nhân tạo: MLP có các siêu tham số đáng chú ý:

- hidden_layer_sizes: Số nút lớp ẩn. Các giá trị được thử nghiệm là [100, 200, 500].
- alpha: hệ số chính quy hoá. Các giá trị được thử nghiệm là [0.0001, 0.001].
- learning_rate: Chiến thuật thay đổi tốc độ học. Các loại chiến thuật được thử nghiệm là ['constant', 'invscaling', 'adaptive'].

C. Cài đặt thực nghiệm

1) Môi trường: Mọi thực nghiệm đều sử dụng Colab notebook, cho phép chạy jupyter notebook miễn phí trên nền web.

2) Thư viện:

- NumPy: NumPy là một thư viện toán học cho phép làm việc hiệu quả với ma trận và mảng với tốc độ xử lý nhanh.
- Pandas: Một thư viện giúp thao tác trên dữ liệu dạng bảng.
- Sklearn: Các mô hình máy học.
- Sklearn.model_selection: Chia dữ liệu thành các tập huấn luyện và kiểm thử.
- Sklearn.metrics: Tính độ chính xác của mô hình.
- Seaborn, Matplotlib: Các thư viện trực quan hoá dữ liệu.
- Imbalanced-learn: Cung cấp thuật toán SMOTE và các biến thể giúp xử lý dữ liệu không cân bằng.

D. Độ đo

1) Độ chính xác: Độ chính xác là tỷ lệ các trường hợp được dự đoán đúng trên tổng số các trường hợp. Độ chính xác giúp đánh giá hiệu quả khả năng dự đoán của mô hình trên một bộ dữ liệu. Độ chính xác càng cao thì mô hình càng tốt. Độ chính xác được tính như sau:

$$Acc. = \frac{Correctprediction}{Totalcases} = \frac{TP + TN}{TP + FP + TN + FP} \quad (2)$$

Với TP, FP, TN, FN lần lượt viết tắt cho true positive, false positive, true negative, false negative. Độ chính xác không được dùng trong bài báo cáo này bởi nó tỏ ra không hiệu quả khi dữ liệu bị mất cân bằng, nhưng nó là tiền đề cho độ đo tiếp theo.

2) Độ chính xác có trọng số: Bởi vì sự mất cân bằng dữ liệu, mô hình trong bài báo cáo này được đánh giá chủ yếu bằng độ chính xác có trọng số thay vì độ chính xác. Độ đo này giống như độ chính xác, trừ việc các điểm dữ liệu ở lớp thiểu số được gán một trọng số lớn hơn lớp đa số, biểu thị rằng một điểm dữ liệu ở lớp này quan trọng hơn một điểm dữ liệu lớp đa số.

3) AUC ROC: AUC-ROC là một thước đo quan trọng trong học máy, đặc biệt là trong bài toán phân loại. Nó đo lường chất lượng dự đoán của mô hình, bất kể ngưỡng phân loại. Về cơ bản, nó định lượng sự cân bằng giữa độ nhạy (tỷ lệ dương tính thực) và độ đặc hiệu (tỷ lệ dương tính giả). Điểm AUC-ROC càng gần 1 thì mô hình càng phân biệt tốt hơn giữa các lớp tích cực và tiêu cực [5]

IV. Thực nghiệm

A. Tinh chỉnh mô hình

Các mô hình được tinh chỉnh siêu tham số bằng hàm GridsearchCV của thư viện sklearn. GridsearchCV (viết tắt cho Grid Search Cross Validation) là một kĩ thuật trong máy học giúp tìm bộ siêu tham số mà mô hình đạt hiệu suất cao nhất. Nó yêu cầu một mô hình đầu vào, các lựa chọn siêu tham số và hàm đánh giá, sau đó tìm kiếm mọi kết hợp có thể mà chọn ra bộ siêu tham số tốt nhất sử dụng kiểm định chéo. Trong bài báo cáo, cv bằng 5 và hàm đánh giá là độ chính xác có trọng số.
Kết quả siêu tham số tối ưu thu được như sau:

- Hồi quy logistic: {'C': 10, 'penalty': 'l2', 'solver': 'newton-cg'}
- K-NN: {'metric': 'manhattan', 'n_neighbors': 3, 'weights': 'distance'}

Bảng I
KẾT QUẢ THỰC NGHIỆM

	Linear	Phân cụm		Cây						SVM	NN
	Logistic Reg.	KNN	NCA+KNN	Dec. Tree	Ran. Forest	CatBoost	LGBM	AdaBoost	XGBoost		MLP
Balanced Acc.	**0.659627**	0.567380	0.622769	0.517577	0.555389	0.598583	0.586388	0.636463	0.583594	0.544216	0.617659
AUC ROC	0.737975	0.584208	0.650225	0.517918	0.728573	**0.744516**	0.709497	0.724486	0.724486	0.690694	0.706227

- Cây quyết định: {'criterion': 'gini', 'max_depth': 10, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 2}
- Random Forest: {'criterion': 'gini', 'max_depth': 10, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2}
- SVM: {'C': 10, 'kernel': 'rbf'}
- MLP: {'alpha': 0.0001, 'hidden_layer_sizes': 200, 'learning_rate': 'constant'}

B. Kết quả thực nghiệm

Kết quả thực nghiệm được thể hiện trong bảng II. Nhận xét rằng, về mặt bằng chung, các mô hình đều cho kết quả tạm được, nhưng thấp so với mức có thể áp dụng mô hình ra thực tế. Có ba mô hình tốt nhất là CatBoost, hồi quy Logistic và AdaBoost, trong đó CatBoost là mô hình có AUC ROC cao nhất (0.744516), cho thấy khả năng phân biệt giữa các lớp tốt nhất Balanced Accuracy cũng ở mức khá (0.598583); Hồi quy Logistic là mô hình có Balanced Accuracy cao nhất (0.659627) và AUC ROC rất tốt (0.737975); AdaBoost là mô hình có cả Balanced Accuracy (0.636463) và AUC ROC (0.724486) đều ở mức cao, cho thấy đây là một mô hình rất cân đối và mạnh mẽ. Về các nhóm, có thể thấy một số mô hình dạng cây và mô hình tuyến tính cho kết quả tốt, trong khi các mô hình nhóm khác chỉ cho kết quả từ không tốt lắm đến tạm được.

V. PHÂN TÍCH LỖI, HƯỚNG PHÁT TRIỂN

Bài báo cáo chọn ra ba mô hình hoạt động tốt để phân tích những dự đoán sai của các mô hình. Ba mô hình mô hình dạng cây được chọn là CatBoost, XGBoost và LGBM. Thực hiện quan sát các đặc điểm của các dự đoán sai, nhóm nhận ra những vấn đề cấp thiết. Đầu tiên, các dự đoán sai thiếu đi những đặc điểm quan trọng. Cụ thể, Các đặc điểm như EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance, OverTime có thể không đủ mạnh để phân biệt rõ ràng các trường hợp. Thêm nữa, các mô hình có thể chưa học được các mối quan hệ ẩn sâu giữa các đặc điểm, hoặc dữ liệu huấn luyện có thể đã gây ra hiện tượng học sai lệch.
Với những thực nghiệm và phân tích trên, nhóm vạch ra những hướng phát triển có thể cho các công trình tiếp theo. Đầu tiên, dữ liệu có những vấn đề nhất định cần phải xem xét như mất cân bằng, nhiều ngoại lai, số mẫu khá ít và dữ liệu này là dữ liệu giả tưởng. Việc có một bộ dữ liệu thật sự với nhiều mẫu dữ liệu hơn, thu nhập đồng đều ở nhiều nơi khác nhau sẽ giúp việc phân tích được thực tế và có ích hơn. Thứ hai, các mô hình được thực nghiệm trong bài báo chưa được chuyên biệt hoá cho bài toán dự đoán thất nghiệp và có kết quả chưa đủ tốt để áp dụng trong thực tế. Các công trình trong tương lai có thể tập trung hơn vào việc xây dựng các mô hình có tính chuyên biệt hoá cao, từ đó cho ra hiệu suất tốt hơn và có thể áp dụng trong thực tế.

VI. KẾT LUẬN

Tỉ lệ nghỉ việc cao gây tổn thất rất lớn cho các doanh nghiệp về mặt tài chính và khả năng phát triển. Bài báo cáo này đã phân tích được một số yếu tố có khả năng ảnh hưởng đến khả năng nghỉ việc của một nhân viên. Thêm nữa, báo báo cáo đã huấn luyện và kiểm tra khả năng của các mô hình máy học trên tập dữ liệu về khả năng nghỉ việc, cho biết khả năng của từng loại mô hình trong bài toán này. Các loại mô hình được thực nghiệm bao gồm năm nhóm: Mô hình tuyến tính, mô hinh phân cụm, mô hình dạng cây, SVM và mạng nơ-ron nhân tạo. Kết quả cho thấy rằng các mô hình hồi quy Logistic, CatBoost và AdaBoost là các mô hình cho kết quả tốt nhất, nhưng kết quả chưa đạt đến mức áp dụng được trong thực tế. Từ kết quả của bài báo cáo, doanh nghiệp có một hướng đi mới nhằm thay đổi, cải thiện để giữ chân nhân tài, đồng thời chuẩn bị cho những mất mát có thể sẽ có trong tương lai gần, tránh được những sự trì trệ không đáng có.

TÀI LIỆU

[1] Anh Le. Thị trường lao động 2023: 75 URL https://nghenghiep.vieclam24h.vn/toa-do-nhan-tai/bao-cao-thi-truong-lao-dong-2023/.

[2] Vương Trần. Gần 19.000 công chức, viên chức thôi việc trong 1 năm qua, Jul 2023. URL https://laodong.vn/thoi-su/gan-19000-cong-chuc-vien-chuc-thoi-viec-trong-1-nam-qua-1218360ldo.

[3] Sarah S. Alduayj and Kashif Rajpoot. Predicting employee attrition using machine learning. *2018 International Conference on Innovations in Information Technology (IIT)*, Nov 2018. doi: 10.1109/innovations.2018.8605976.

[4] Alberto Cabrera. *"Logistic Regression Analysis in Higher Education: An Applied Perspective*, volume 10, pages 225–256. 01 1994.

[5] Sep 2023. URL https://funix.edu.vn/chia-se-kien-thuc/auc-roc-trong-xay-dung-cac-mo-hinh-ai-hieu-suat-cao.