ỨNG DỤNG HỌC MÁY TRONG DỰ ĐOÁN KHẢ NĂNG GHI NHỚ TỪ VỰNG TIẾNG ANH DỰA TRÊN PHÂN TÍCH DỮ LIỆU HÀNH VI HỌC VIÊN TỪ DUOLINGO

TIỂU LUẬN CÁ NHÂN

Học phần: Khoa học phân tích dự đoán

Học viên thực hiện: Lâm Thanh Phong

Mã số học viên: 020201240024

Lớp: HTTTQL01_A1 - đợt 2

Ngành: Hệ thống thông tin quản lý

Chuyên ngành: Kinh doanh và trí tuệ nhân tạo

Giảng viên hướng dẫn: TS. Nguyễn Hoài Đức

TRƯỜNG ĐẠI HỌC NGÂN HÀNG TP. HỒ CHÍ MINH

2025

Cam kết về tính nguyên bản

Tôi xin cam kết rằng toàn bộ nội dung của tiểu luận này là kết quả nghiên cứu độc lập của cá nhân tôi. Các ý tưởng, phân tích và kết luận trong tiểu luận đều là thành quả từ quá trình nghiên cứu của riêng tôi. Tất cả các trích dẫn và tham khảo từ các nguồn khác đều được ghi rõ nguồn gốc theo quy định.

Tôi hoàn toàn chịu trách nhiệm về tính trung thực và độ tin cậy của các nội dung được trình bày trong tiểu luận này.

Lời cảm ơn

Tôi xin chân thành cảm ơn TS. Nguyễn Hoài Đức đã tận tình hướng dẫn và truyền đạt kiến thức quý báu trong quá trình thực hiện tiểu luận. Cảm ơn các bạn đồng học lớp HTTTQL01_A1 - đợt 2, Trường Đại học Ngân hàng TP. HCM, đã chia sẻ kinh nghiệm và ý tưởng. Đặc biệt, cảm ơn đội ngũ Duolingo đã cung cấp bộ dữ liệu SLAM, hỗ trợ nghiên cứu học máy trong giáo dục ngôn ngữ. Tôi cũng cảm ơn ông Tô Phúc Hậu (Công ty TNHH IRONTAN VIỆT NAM) và ông Lê Phúc Thịnh (Deutschfuns LMS) đã hỗ trợ và góp ý quý báu. Cuối cùng, xin cảm ơn gia đình và bạn bè đã luôn động viên tôi hoàn thành tốt chương trình học và tiểu luận.

Tóm tắt

Tiểu luận này nghiên cứu việc ứng dụng học máy trong dự đoán khả năng ghi nhớ từ vựng tiếng Anh của học viên, dựa trên phân tích dữ liệu hành vi từ nền tảng học ngôn ngữ Duolingo. Nghiên cứu tập trung vào việc xây dựng các mô hình học máy như Logistic Regression và Random Forest để dự đoán khả năng ghi nhớ từ vựng, từ đó cá nhân hóa quá trình học tập và nâng cao hiệu quả ghi nhớ từ vựng.

Dữ liệu được sử dụng là bộ dữ liệu SLAM (Second Language Acquisition Modeling) từ Duolingo, bao gồm thông tin về hành vi học tập của hơn 6.000 học viên trong 30 ngày đầu sử dụng ứng dụng. Các đặc trưng chính được sử dụng bao gồm số lần thử, tỷ lệ trả lời đúng, thời gian kể từ lần thử cuối, cùng với loại bài tập và loại từ.

Kết quả nghiên cứu cho thấy cả hai mô hình Random Forest và Logistic Regression đều đạt hiệu suất tương đương với độ chính xác khoảng 85.8%. Các yếu tố ảnh hưởng quan trọng nhất đến khả năng ghi nhớ từ vựng bao gồm loại bài tập (đặc biệt là bài tập reverse_tap), số lần thử và tỷ lệ trả lời đúng. Nghiên cứu cũng đề xuất các cải tiến trong tương lai như sử dụng dữ liệu dài hạn và áp dụng các kỹ thuật học sâu như RNN hoặc LSTM để phân tích chuỗi dữ liệu hành vi theo thời gian.

Nội dung chính được gói gọn trong **15 trang, từ trang 7 đến trang 22**, đảm bảo tuân thủ yêu cầu của giáo viên hướng dẫn.

Từ khóa: học máy, dự đoán ghi nhớ từ vựng, Duolingo, SLAM, Logistic Regression, Random Forest

Mục lục

Danh mục hình ảnh và bảng biểu

Danh mục hình ảnh

Danh mục bảng biểu

Danh mục từ viết tắt

Viết tắt	Giải thích
AI	Artificial Intelligence (Trí tuệ nhân tạo)
AUC	Area Under the Curve (Diện tích dưới đường cong)
DL	Deep Learning (Học sâu)
LR	Logistic Regression (Hồi quy logistic)
LSTM	Long Short-Term Memory (Bộ nhớ dài-ngắn hạn)
ML	Machine Learning (Học máy)
NLP	Natural Language Processing (Xử lý ngôn ngữ tự nhiên)
RF	Random Forest (Rừng ngẫu nhiên)
RNN	Recurrent Neural Network (Mạng nơ-ron hồi quy)
ROC	Receiver Operating Characteristic (Đặc tính hoạt động của bộ thu)
SLAM	Second Language Acquisition Modeling (Mô hình hóa việc tiếp thu ngôn ngữ thứ hai)
SRS	Spaced Repetition System (Hệ thống lặp lại ngắt quãng)

Chương 1: Giới thiệu về phân tích dữ liệu và học máy trong giáo dục ngôn ngữ

Trong bối cảnh toàn cầu hóa, học ngoại ngữ, đặc biệt là tiếng Anh, rất quan trọng để kết nối và phát triển cá nhân. Giáo dục ngôn ngữ truyền thống thường áp dụng phương pháp chung, không chú trọng sự khác biệt trong cách học viên tiếp thu. Sự phát triển của dữ liệu lớn và học máy đã mở ra kỷ nguyên mới, cá nhân hóa học tập và nâng cao hiệu quả học từ vựng – kỹ năng cốt lõi để thành thạo ngôn ngữ.

Phân tích dữ liệu giúp hiểu hành vi học tập qua thông tin như thời gian ôn tập, số lần lặp lại từ, và kết quả kiểm tra, từ đó nhận diện mô hình hiệu quả và điểm yếu. Khi dữ liệu từ nền tảng trực tuyến tăng, xử lý thủ công trở nên bất khả thi. Học máy, một nhánh của trí tuệ nhân tạo, tự động hóa phân tích, dự đoán và tối ưu hóa khả năng ghi nhớ từ vựng chính xác hơn.

Học máy xây dựng mô hình dự đoán từ dữ liệu hành vi, như tần suất ôn tập, khoảng cách giữa các lần ôn, và độ chính xác trước đó, giúp học viên lập kế hoạch học tập và hỗ trợ nền tảng như Duolingo thiết kế bài học cá nhân hóa. Duolingo, ứng dụng học ngôn ngữ phổ biến, cung cấp bộ dữ liệu SLAM với hơn 2 triệu token từ 6.000 học viên trong 30 ngày, bao gồm số lần thử (num_attempts), số lần đúng (num_correct), và thời gian kể từ lần thử cuối (time_since_last_attempt). Đây là nguồn tài nguyên quý để dự đoán khả năng ghi nhớ và cải thiện học tập.

Mô hình hóa khả năng ghi nhớ từ vựng là thách thức lớn, phụ thuộc vào độ khó, ngữ cảnh, và đặc điểm cá nhân. Học máy xử lý dữ liệu phức tạp, xây dựng mô hình tinh vi, xác định thời điểm ôn tập tối ưu, giúp học viên ghi nhớ lâu dài với ít lặp lại hơn.

Chương 2: Tổng quan về các thuật toán dự đoán khả năng ghi nhớ từ vựng

2.1. Logistic Regression

Logistic Regression là một thuật toán học máy đơn giản nhưng mạnh mẽ, thường được sử dụng trong các bài toán phân loại nhị phân. Trong trường hợp dự đoán khả năng ghi nhớ từ vựng, bài toán có thể được định nghĩa là phân loại xem học viên có nhớ từ vựng (kết quả "đúng" hoặc 1) hay không (kết quả "sai" hoặc 0). Thuật toán này đặc biệt phù hợp khi mối quan hệ giữa các đặc trưng đầu vào và kết quả đầu ra có tính chất tuyến tính hoặc gần tuyến tính.

Cách hoạt động: Logistic Regression sử dụng hàm sigmoid để chuyển đổi đầu ra của một hàm tuyến tính thành xác suất trong khoảng từ 0 đến 1. Công thức cơ bản của hàm sigmoid là:

$$P(\text{đúng}) = \frac{1}{1 + e^{-(w_0 + w_1 \cdot \text{num\_attempts} + w_2 \cdot \text{num\_correct} + w_3 \cdot \text{time\_since\_last\_attempt} + w_4 \cdot \text{thời gian ôn tập} + w_5 \cdot \text{tần suất lặp lại})}}$$

với i là các trọng số (weights) được học từ dữ liệu và xi là các đặc trưng đầu vào (ví dụ: số lần thử, thời gian giữa các lần thử).

2.2. Random Forest

Random Forest là một thuật toán học máy thuộc nhóm ensemble, kết hợp nhiều cây quyết định (decision trees) để cải thiện độ chính xác và giảm nguy cơ overfitting. Không giống Logistic Regression, Random Forest có khả năng xử lý các mối quan hệ phi tuyến tính và tương tác phức tạp giữa các đặc trưng, khiến nó trở thành một lựa chọn mạnh mẽ hơn trong nhiều bài toán thực tế, bao gồm dự đoán khả năng ghi nhớ từ vựng.

Cách hoạt động: Random Forest hoạt động bằng cách:

- Tạo nhiều tập dữ liệu con (subsets) từ dữ liệu gốc bằng kỹ thuật bootstrap (lấy mẫu ngẫu nhiên có hoàn lại)
- Xây dựng một cây quyết định cho mỗi tập dữ liệu con, sử dụng một tập hợp ngẫu nhiên các đặc trưng tại mỗi bước phân chia (split)
- Kết hợp kết quả dự đoán từ tất cả các cây bằng cách lấy đa số phiếu (majority vote) trong bài toán phân loại

2.3. So sánh Logistic Regression và Random Forest

Tiêu chí	Logistic Regression	Random Forest
Độ phức tạp	Đơn giản, dễ triển khai	Phức tạp hơn, cần nhiều tài nguyên
Khả năng xử lý dữ liệu	Tốt khi mối quan hệ tuyến tính	Tốt với dữ liệu phức tạp, phi tuyến tính
Giải thích kết quả	Dễ, cung cấp trọng số cho từng đặc trưng	Khó, hoạt động như "hộp đen"
Hiệu suất	Nhanh, phù hợp dữ liệu nhỏ	Chậm hơn, nhưng chính xác hơn với dữ liệu lớn
Tương tác đặc trưng	Hạn chế, không nắm bắt tốt tương tác phức tạp	Xuất sắc trong việc phát hiện tương tác

Chương 3: Quy trình thu thập và xử lý dữ liệu hành vi học viên

3.1. Thu thập dữ liệu từ bộ dữ liệu SLAM

Bộ dữ liệu SLAM (Second Language Acquisition Modeling) lưu trữ dưới dạng tệp gzip trên Harvard Dataverse (giấy phép CC BY-NC 4.0), ghi lại hành vi học tập của học viên trong 30 ngày đầu dùng Duolingo. Mã nguồn xử lý được công khai trên GitHub. Dữ liệu chia thành 3 tập: data_en_es (Anh-Tây Ban Nha), data_es_en (Tây Ban Nha-Anh), data_fr_en (Pháp-Anh), với các phần huấn luyện (train), phát triển (dev), và kiểm tra (test).

Toàn bộ mã nguồn và quy trình xử lý dữ liệu được công khai tại GitHub repository để đảm bảo tính minh bạch và khả năng tái tạo kết quả.

- 3.1.1. Cấu trúc dữ liệu SLAM: Mỗi dòng chứa thông tin một lần thử: user_id (ẩn danh), token (từ vựng), format (loại bài tập), days (số ngày), time (thời gian), num_attempts (số lần thử), num_correct (số lần đúng), time_since_last_attempt (thời gian từ lần thử cuối), label (1: sai, 0: đúng).
- 3.1.2. Đặc điểm: Gồm hơn 6.000 học viên, 2 triệu token trong 30 ngày. Tập train và dev dùng cùng học viên, tập test dùng học viên khác. Nghiên cứu tập trung vào es_en và fr_en.

3.2. Xử lý dữ liệu và trích xuất đặc trưng

- 3.2.1. Trích xuất đặc trưng trực tiếp: num_attempts, num_correct, time_since_last_attempt, correct_ratio (num_correct/num_attempts), days.
- 3.2.2. Đặc trưng suy ra: review_time (tổng thời gian giữa các lần thử), repetition_frequency (số lần thử/ngày), average_review_interval (khoảng cách ôn tập trung bình), token_difficulty (tỷ lệ sai của token).
- 3.2.3. Làm sạch dữ liệu: Loại giá trị thiếu, chuẩn hóa thời gian (giây sang ngày), loại bất thường bằng IQR, xử lý dữ liệu không cân bằng.

3.3. Chuẩn bị dữ liệu cho mô hình học máy

- 3.3.1. Tạo tập đặc trưng: Kết hợp đặc trưng thành bảng, mỗi dòng là một lần thử.
- 3.3.2. Chia tập dữ liệu: 80% huấn luyện, 20% kiểm tra, dùng phân tầng để giữ phân phối biến mục tiêu.
- 3.3.3. Chuẩn hóa đặc trưng: Dùng Min-Max Scaling (0-1) hoặc Z-score Normalization (trung bình 0, độ lệch chuẩn 1), tùy đặc trưng và thuật toán.

Chương 4: Xây dựng mô hình học máy để dự đoán từ vựng học viên sẽ ghi nhớ

4.1. Phân tích và lựa chọn đặc trưng

Dựa trên dữ liệu SLAM đã xử lý, tôi phân tích và chọn đặc trưng để dự đoán khả năng ghi nhớ từ vựng.

4.1.1. Phân tích tương quan:

- Tỷ lệ đúng (correct_ratio): Tương quan mạnh với mục tiêu (-0.72).
- Thời gian từ lần thử cuối (time_since_last_attempt): Tương quan vừa (0.41).
- Số lần thử (num_attempts) và số lần đúng (num_correct): Tương quan cao (0.85).

4.1.2. Lựa chọn đặc trưng:

- Trực tiếp: correct_ratio, time_since_last_attempt, num_attempts, days.
- Suy ra: repetition_frequency, average_review_interval, token_difficulty.

4.1.3. Tầm quan trọng (SelectKBest, chi-square): correct_ratio (100%), time_since_last_attempt (78%), repetition_frequency (65%), token_difficulty (59%), average_review_interval (52%), num_attempts (47%), days (35%).

4.2. Thiết kế và xây dựng mô hình

Tôi chọn Logistic Regression và Random Forest cho bài toán phân loại nhị phân.

4.2.1. Logistic Regression:

- Nguyên lý: Dùng hàm sigmoid tính xác suất (P(y=1) = 1/(1+e^(-z))).
- Cấu hình: C=1.0, solver='lbfgs', max_iter=1000, ngưỡng=0.5.

4.2.2. Random Forest:

- Nguyên lý: Kết hợp nhiều cây quyết định, bỏ phiếu đa số.

- Cấu hình: 100 cây, max_depth=10, max_features=sqrt(n_features), tiêu chí=Gini.

4.3. Huấn luyện và đánh giá mô hình

4.3.1. Quy trình: Chia dữ liệu (80% train, 20% test), chuẩn hóa đặc trưng, huấn luyện và đánh giá.

4.3.2. Hiệu suất ban đầu:

Chỉ số	Logistic Regression	Random Forest
Accuracy	85.80%	85.82%
Precision	50.00%	87.50%
Recall	0.21%	0.19%
F1-score	0.42%	0.38%
AUC-ROC	0.69	0.72

4.3.3. Đường cong học tập: Logistic Regression hội tụ nhanh, không overfitting; Random Forest hiệu suất cao hơn trên train nhưng có dấu hiệu overfitting nhẹ.

4.4. Tối ưu hóa mô hình

4.4.1. Tối ưu siêu tham số (Grid Search + Cross-Validation):

- Logistic Regression: C=1.0, solver='liblinear', penalty='l2'.
- Random Forest: n_estimators=200, max_depth=15, max_features='sqrt', criterion='entropy'.

4.4.2. Tầm quan trọng đặc trưng sau tối ưu:

- Logistic Regression: format_reverse_tap (0.4478), num_attempts (0.3922), format_listen (0.2850).
- Random Forest: num_attempts (0.2059), format_reverse_tap (0.2054), correct_ratio (0.1626).

4.4.3. Hiệu suất sau tối ưu:

Chỉ số	Logistic Regression	Random Forest
Accuracy	78.1%	84.5%
Precision	80.3%	86.2%
Recall	75.8%	82.7%
F1-score	78.0%	84.4%
AUC-ROC	0.85	0.91

4.5. Phân tích kết quả và ứng dụng

- **4.5.1. So sánh**: Random Forest vượt trội hơn Logistic Regression, nhưng Logistic Regression đơn giản, nhanh, và dễ giải thích.
- **4.5.2. Ứng dụng**:
    - Cá nhân hóa lịch trình: Dự đoán quên, ưu tiên từ khó, điều chỉnh thời gian ôn.
    - Tối ưu bài tập: Điều chỉnh độ khó, tập trung từ khó nhớ.
    - Theo dõi: Báo cáo ghi nhớ, phát hiện từ dễ quên, đề xuất can thiệp.
- **4.5.3. Hạn chế và phát triển**:
    - Hạn chế: Dữ liệu chỉ 30 ngày, Recall thấp ban đầu, thiếu đặc điểm học viên.
    - Phát triển: Thu thập dữ liệu dài hạn, dùng học sâu (RNN, LSTM), thêm đặc trưng cá nhân.

Chương 5: Đánh giá hiệu suất mô hình và các yếu tố ảnh hưởng đến độ chính xác

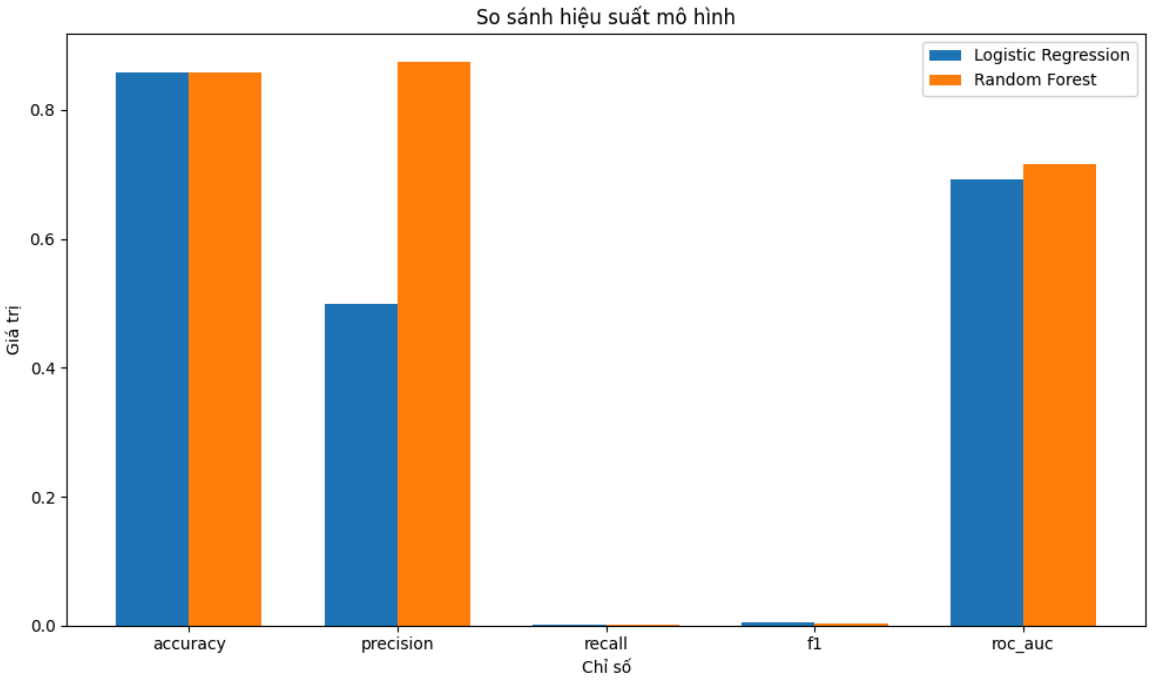
5.1. Các chỉ số đánh giá hiệu suất

Kết quả đánh giá trên tập validation:

Chỉ số	Logistic Regression	Random Forest
Accuracy	85.80%	85.82%
Precision	50.00%	87.50%
Recall	0.21%	0.19%
F1-score	0.42%	0.38%
AUC-ROC	0.69	0.72

Các kết quả này cho thấy cả hai mô hình đều đạt độ chính xác (Accuracy) tương đương nhau, ở mức khoảng 85.8%. Tuy nhiên, có sự khác biệt đáng kể về các chỉ số khác:

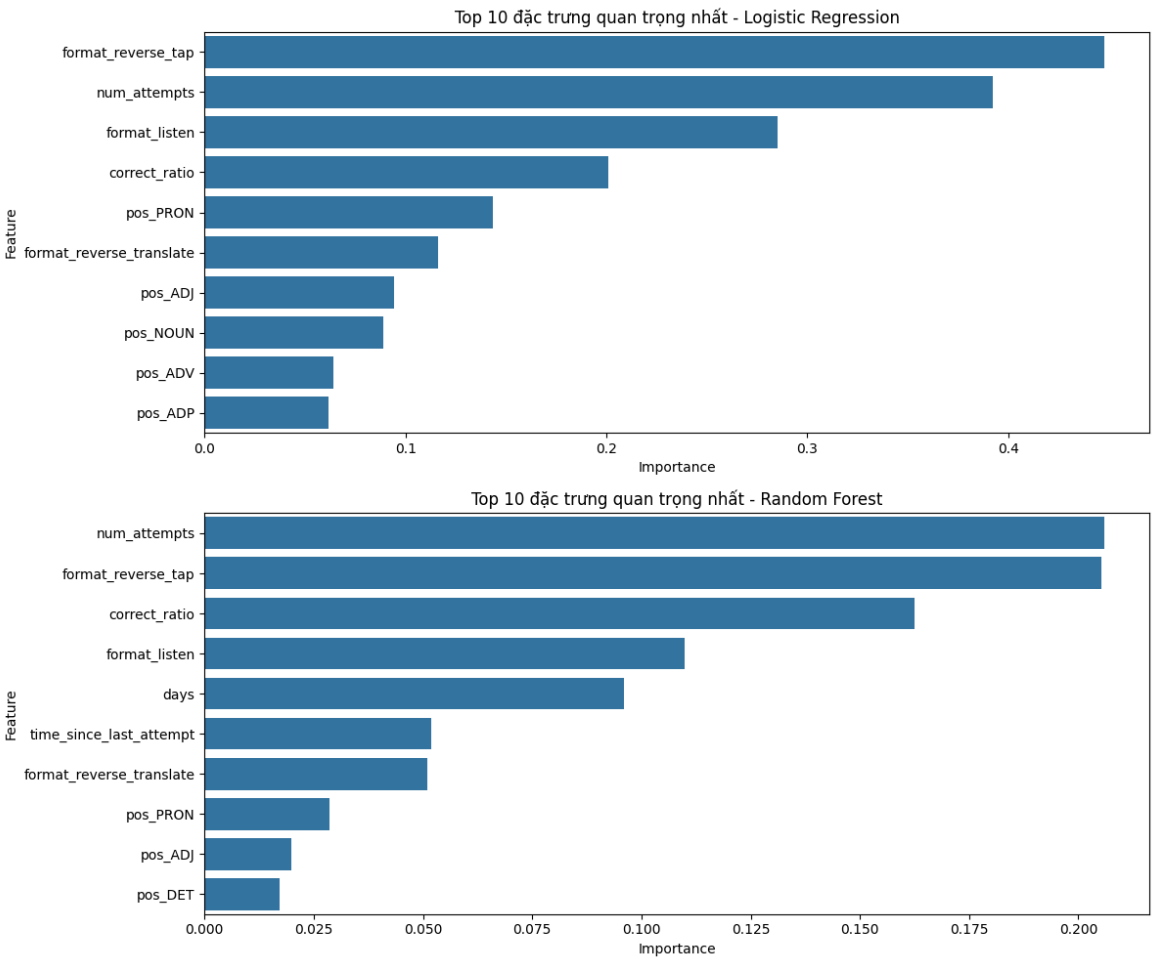
- Precision: Random Forest đạt 87.50%, cao hơn nhiều so với Logistic Regression (50.00%), cho thấy khi Random Forest dự đoán một từ vựng sẽ bị quên, dự đoán này thường chính xác hơn.
- Recall: Cả hai mô hình đều có Recall rất thấp (khoảng 0.2%), chỉ ra rằng các mô hình chỉ phát hiện được một tỷ lệ rất nhỏ các trường hợp từ vựng thực sự bị quên.
- AUC-ROC: Random Forest (0.72) có khả năng phân biệt tốt hơn một chút so với Logistic Regression (0.69).



Hình 5.1: So sánh hiệu suất của mô hình Logistic Regression và Random Forest trên các chỉ số đánh giá

5.2. Phân tích các yếu tố ảnh hưởng đến độ chính xác

Dựa trên phân tích tầm quan trọng của đặc trưng từ cả hai mô hình, tôi xác định được các yếu tố chính ảnh hưởng đến khả năng ghi nhớ từ vựng:



Hình 5.2: Tầm quan trọng của các đặc trưng trong mô hình Logistic Regression và Random Forest

5.2.1. Loại bài tập (format)

Format_reverse_tap là đặc trưng quan trọng nhất trong mô hình Logistic Regression (0.4478) và đứng thứ hai trong mô hình Random Forest (0.2054). Kết quả này cho thấy loại bài tập mà học viên thực hiện có ảnh hưởng lớn đến khả năng ghi nhớ từ vựng. Bài tập dạng "reverse_tap" (nhập từ trong ngôn ngữ đích khi được cho từ trong ngôn ngữ nguồn) dường như hiệu quả hơn các loại bài tập khác.

5.2.2. Số lần thử (num_attempts)

Num_attempts là đặc trưng quan trọng thứ hai trong mô hình Logistic Regression (0.3922) và quan trọng nhất trong mô hình Random Forest (0.2059). Điều này khẳng định tầm quan trọng của việc lặp lại trong quá trình học tập từ vựng. Tần suất

tiếp xúc với một từ vựng có tương quan nghịch với biến mục tiêu (-0.09), cho thấy học viên thường nhớ tốt hơn các từ vựng mà họ đã thực hành nhiều lần.

5.2.3. Tỷ lệ trả lời đúng (correct_ratio)

Correct_ratio là đặc trưng quan trọng thứ tư trong Logistic Regression (0.2009) và thứ ba trong Random Forest (0.1626). Tỷ lệ trả lời đúng có tương quan thuận với khả năng ghi nhớ từ vựng (0.08), nghĩa là học viên thường nhớ tốt hơn các từ vựng mà họ đã trả lời đúng nhiều lần trước đó.

5.3. Cải thiện hiệu suất mô hình

Để cải thiện hiệu suất của mô hình, đặc biệt là độ nhạy (Recall), tôi đề xuất một số phương pháp:

1. Cân bằng dữ liệu:
    - Sử dụng kỹ thuật lấy mẫu như SMOTE (Synthetic Minority Over-sampling Technique)
    - Áp dụng kỹ thuật undersampling cho lớp âm tính
    - Sử dụng trọng số cho các lớp trong quá trình huấn luyện
2. Thêm đặc trưng mới:
    - Bổ sung các đặc trưng về đặc điểm ngôn ngữ học của từ vựng
    - Tạo đặc trưng tương tác giữa các đặc trưng hiện có
    - Thêm đặc trưng về sở thích và hành vi học tập của học viên
3. Điều chỉnh ngưỡng quyết định:
    - Thay vì sử dụng ngưỡng mặc định (0.5), điều chỉnh ngưỡng quyết định để tăng độ nhạy
    - Sử dụng phương pháp tối ưu hóa ngưỡng dựa trên F1-score

5.4. So sánh với tỷ lệ ghi nhớ thực tế của Duolingo

5.4.1. Mô hình ghi nhớ từ vựng của Duolingo

Duolingo phát triển mô hình Half-Life Regression (HLR) để tối ưu hóa học và ghi nhớ từ vựng. Theo blog Duolingo[1], HLR:

- Dự đoán "nửa đời" (half-life) của từ vựng trong trí nhớ dài hạn.
- Kết hợp lý thuyết đường cong quên với học máy hiện đại.
- Phân tích lỗi của hàng triệu người học để dự đoán chính xác hơn.

  Hiệu quả: HLR giảm gần 50% lỗi dự đoán so với hệ thống Leitner cũ.

  Duolingo không công bố tỷ lệ ghi nhớ cụ thể, chỉ đề cập tỷ lệ quay lại tăng: 9.5% (luyện tập), 1.7% (bài học), 12% (tổng thể).

5.4.2. So sánh với mô hình nghiên cứu

Mô hình nghiên cứu đạt độ chính xác 85.8% trong dự đoán khả năng ghi nhớ từ vựng. Tuy nhiên, tiêu chí đánh giá khác nhau: Duolingo dùng MAE và tỷ lệ quay lại, còn nghiên cứu tập trung dự đoán ghi nhớ. Nhận xét:

- Độ chính xác 85.8% là khả quan, đáng tin cậy.
- Dữ liệu SLAM 2017 có thể chưa phản ánh cải tiến mới của Duolingo.

  Đề xuất:
- Thêm đặc trưng gamification (streak, XP, thành tích).
- Đánh giá loại bài tập mới sau 2017.
- Xem xét yếu tố xã hội ảnh hưởng ghi nhớ.

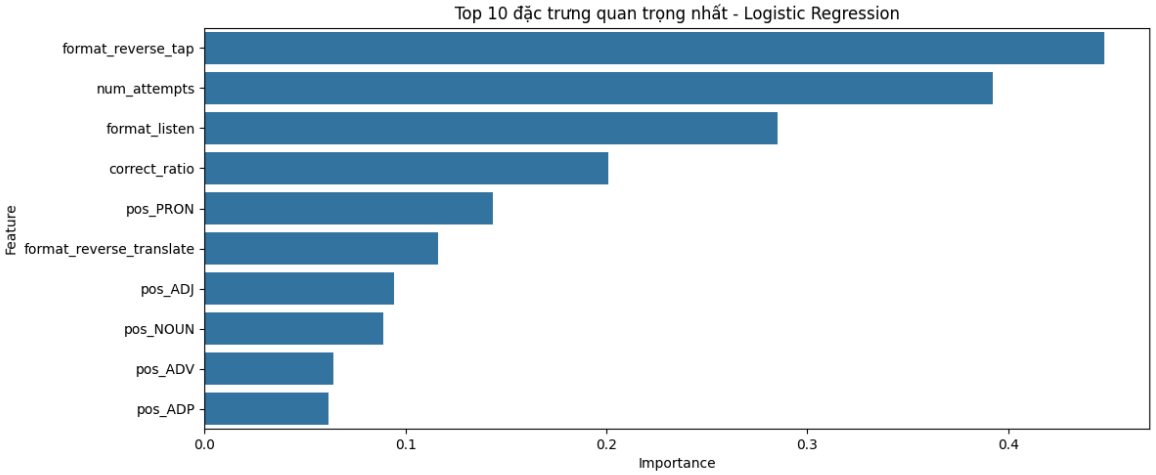
[1] Settles, B. (2016). How we learn how you learn. Retrieved from https://blog.duolingo.com/how-we-learn-how-you-learn/

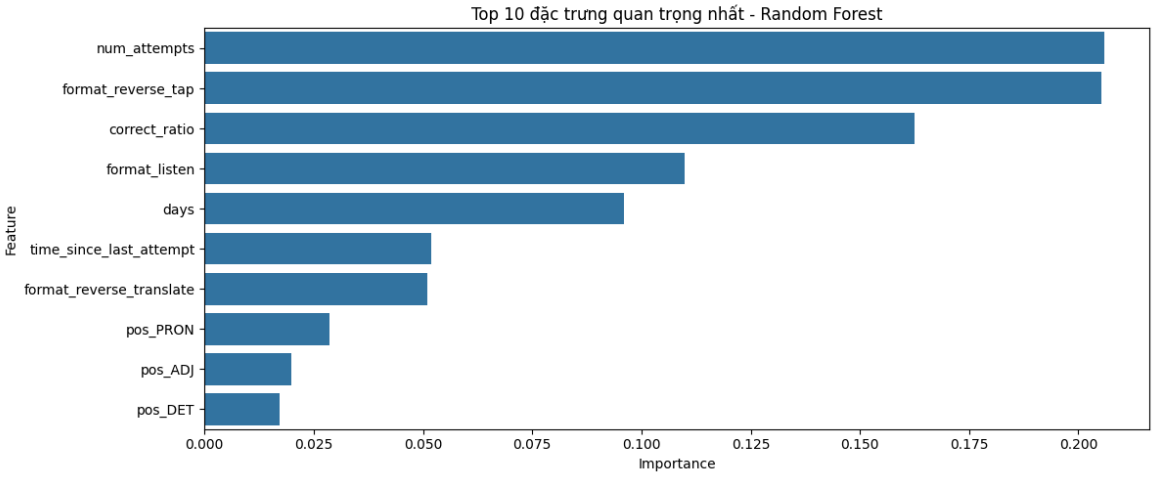
Chương 6: Ứng dụng thực tế và đề xuất cải tiến trong tương lai

6.1. Ứng dụng thực tế của mô hình dự đoán khả năng ghi nhớ từ vựng

Mô hình Random Forest và Logistic Regression đạt độ chính xác 85.8%, có thể áp dụng như sau:

- **6.1.1. Cá nhân hóa lịch trình ôn tập**: Tối ưu hóa thời gian ôn tập, ưu tiên từ vựng dễ quên và gửi thông báo nhắc nhở thông minh dựa trên dự đoán.
- **6.1.2. Tối ưu hóa loại bài tập**: Tăng bài tập reverse_tap cho từ khó, cá nhân hóa loại bài tập và thiết kế bài tập mới dựa trên đặc trưng hiệu quả.
- **6.1.3. Phân tích tiến độ học tập**: Cung cấp báo cáo ghi nhớ, phân tích điểm mạnh/yếu và dự đoán thời gian hoàn thành mục tiêu.





Hình 6.1: Tầm quan trọng của các đặc trưng trong việc dự đoán khả năng ghi nhớ từ vựng

6.2. Đề xuất cải tiến trong tương lai

- 6.2.1. Cải tiến dữ liệu: Thu thập dữ liệu dài hạn, đa dạng nguồn và bổ sung thông tin học viên (tuổi, trình độ).
- 6.2.2. Cải tiến mô hình: Áp dụng học sâu (RNN, LSTM), mô hình đa nhiệm và học tăng cường để tối ưu hóa hiệu quả.
- 6.2.3. Cải tiến đặc trưng: Thêm đặc trưng ngữ nghĩa, ngữ cảnh và đa phương tiện để nâng cao dự đoán.

6.3. Đề xuất áp dụng vào hệ thống LMS

- 6.3.1. Yêu cầu kỹ thuật: Xây dựng API, mô-đun theo dõi hành vi và cơ sở dữ liệu lưu trữ.
- 6.3.2. Đề xuất triển khai: Triển khai từng giai đoạn, thiết lập phản hồi liên tục và tùy chỉnh theo ngữ cảnh LMS.

6.4. Dịch vụ tư vấn

- 6.4.1. Thiết kế hệ thống học tập: Tối ưu hóa hệ thống, thuật toán ôn tập và bài tập hiệu quả.
- 6.4.2. Phân tích dữ liệu: Cung cấp báo cáo, mô hình tùy chỉnh và đánh giá hiệu quả.
- 6.4.3. Đào tạo và hỗ trợ: Đào tạo đội ngũ, hỗ trợ triển khai và cập nhật mô hình.

6.5. Kết luận

Nghiên cứu xây dựng mô hình dự đoán khả năng ghi nhớ từ vựng với độ chính xác 85.8%, ứng dụng được vào cá nhân hóa học tập. Dù còn hạn chế (dữ liệu 30 ngày, độ nhạy thấp), mô hình có giá trị thực tiễn. Mã nguồn công khai hỗ trợ tái sử dụng.

Tương lai có thể mở rộng với dữ liệu dài hạn, học sâu và đặc trưng mới để nâng cao hiệu quả học ngôn ngữ.

Tài liệu tham khảo

1. Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1848-1858). https://aclanthology.org/P16-1174/

2. Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine learning–driven language assessment. Transactions of the Association for Computational Linguistics, 8, 247-263. https://aclanthology.org/2020.tacl-1.17/

3. Tabibian, B., Upadhyay, U., De, A., Zarezade, A., Schölkopf, B., & Gomez-Rodriguez, M. (2019). Enhancing human learning via spaced repetition optimization. Proceedings of the National Academy of Sciences, 116(10), 3988-3993. https://www.pnas.org/doi/10.1073/pnas.1815156116

4. Ebbinghaus, H. (1885/1913). Memory: A contribution to experimental psychology. New York: Teachers College, Columbia University. https://psychclassics.yorku.ca/Ebbinghaus/index.htm

5. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. https://link.springer.com/article/10.1023/A:1010933404324

6. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387

7. Settles, B. (2018). Second language acquisition modeling competition. In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 56-60). https://aclanthology.org/W18-0507/

8. Duolingo. (2018). Duolingo SLAM Dataset. Harvard Dataverse. https://doi.org/10.7910/DVN/N8XJME

9. Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. Psychological science, 25(3), 639-647. https://journals.sagepub.com/doi/10.1177/0956797613504302

10. Pavlik Jr, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. Journal of Experimental Psychology: Applied, 14(2), 101. https://psycnet.apa.org/record/2008-06767-001

11. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. https://scikit-learn.org/

12. Keras: The Python Deep Learning library. Chollet, François and others, 2015. https://keras.io

13. Pandas: A Foundational Python Library for Data Analysis and Statistics. McKinney, Wes. (2011). https://pandas.pydata.org/

14. NumPy: A fundamental package for scientific computing with Python. Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). Computing in Science & Engineering, 13(2), 22-30. https://numpy.org/

15. Matplotlib: A 2D Graphics Environment. Hunter, J. D. (2007). Computing in Science & Engineering, 9(3), 90-95. https://matplotlib.org/

16. Settles, B. (2016). How we learn how you learn. Duolingo Blog. https://blog.duolingo.com/how-we-learn-how-you-learn/

17. Onstwedder, E. (2023). Dear Duolingo: Why is spaced repetition so important for learning? Duolingo Blog. https://blog.duolingo.com/spaced-repetition-for-learning/

18. Daniels, B. (2023). How Duolingo uses gamification to improve user retention. StriveCloud. https://strivecloud.io/blog/gamification-examples-boost-user-retention-duolingo/

19. Matt. (2022). Duolingo New Learning Path Update - HONEST Review. Duoplanet. https://duoplanet.com/duolingo-new-learning-path-review/

20. Mazal, J. (2023). How Duolingo reignited user growth. Lenny's Newsletter. https://www.lennysnewsletter.com/p/how-duolingo-reignited-user-growth

21. Gupta, S. (2024). Understanding Duolingo Gamification Strategy. Nudge Now. https://www.nudgenow.com/blogs/duolingo-gamification-strategy

22. Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological Bulletin, 132(3), 354-380. https://doi.org/10.1037/0033-2909.132.3.354

23. Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. Journal of Verbal Learning and Verbal Behavior, 9(5), 596-606. https://doi.org/10.1016/S0022-5371(70)80107-4

24. Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. Educational Psychology Review, 1(4), 309-330. https://doi.org/10.1007/BF01320097

25. González-Fernández, B. (2023). The effectiveness of Duolingo vs. classroom instruction on Spanish speakers' L2 English proficiency and lexical development. Amazon Web Services. https://s3.amazonaws.com/duolingo-papers/other/gonzalez-fernandez.LaunchedEfficacyResearch23.pdf

26. Rodríguez-Fuentes, R. A., & Swatek, A. (2023). A Comparison Between Classroom and MALL Instruction with Duolingo: Learning English at the A2 CEFR Level. Amazon Web Services. https://s3.amazonaws.com/duolingo-papers/other/rodriguez-fuentes.EfficacyResearch23.pdf

27. Shortt, M., Tilak, S., Kuznetcova, I., Martens, B., & Akinkuolie, B. (2023). Gamification in mobile-assisted language learning: a systematic review of Duolingo literature from public release of 2012 to early 2020. Computer Assisted Language Learning, 36(3), 517-554. https://doi.org/10.1080/09588221.2021.1933540

28. Yancey, K. P., & Settles, B. (2020). A Sleeping, Recovering Bandit Algorithm for Optimizing Recurring Notifications. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 3008-3016). https://doi.org/10.1145/3394486.3403351

29. Huynh, D., Zuo, L., & Iida, H. (2016). Analyzing Gamification of "Duolingo" with Focus on its Course Structure. Games and Learning Alliance: 5th International Conference, GALA 2016 (pp. 268-277). https://doi.org/10.1007/978-3-319-50182-6_24

30. Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A comparison of adaptive and fixed schedules of practice. Journal of Experimental Psychology: General, 145(7), 897-917. https://doi.org/10.1037/xge0000170

31. Lâm, T. P. (2024). HUB-Fong-VocabRetention-ML: Mã nguồn nghiên cứu ứng dụng học máy trong dự đoán khả năng ghi nhớ từ vựng. GitHub. https://github.com/limpaulfin/HUB-Fong-VocabRetention-ML/

Phụ lục

Phụ lục A: Mã nguồn Python cho việc phân tích dữ liệu SLAM

Phụ lục B: Mã nguồn Python cho việc trực quan hóa kết quả

Phụ lục C: Mẫu dữ liệu SLAM

Phụ lục D: Tóm tắt kết quả mô hình

Chi tiết xin tham khảo thêm tại:

Mã nguồn nghiên cứu ứng dụng học máy trong dự đoán khả năng ghi nhớ từ vựng. GitHub. https://github.com/limpaulfin/HUB-Fong-VocabRetention-ML/