TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT ĐHQG -TPHCM

ĐỒ ÁN CUỐI KỲ - HK2 (2022-2023)

MÔN HỌC: PHÂN TÍCH DỮ LIỆU CƠ BẢN

PHÂN TÍCH VÀ TRỰC QUAN HOÁ DỮ LIỆU HỒ SƠ VAY VỐN ĐỂ DỰ ĐOÁN KHẢ NĂNG VỠ NỢ CỦA KHÁCH HÀNG

GIẢNG VIÊN HƯỚNG DẪN: TS. Nguyễn Thôn Dã
MÃ HỌC PHẦN: 222MI21
NHÓM SINH VIÊN THỰC HIỆN:

STT	Họ và Tên	MSSV	Gmail
1	Nguyễn Quốc Huy	K214142068	huynq21414@st.uel.edu.vn
2	Hồ Thị Khánh Ly	K214142071	lyhtk21414@st.uel.edu.vn
3	Ngô Võ Ngọc Mai	K214142072	mainvn21414@st.uel.edu.vn
4	Nguyễn Mai Phương	K214142080	phuongnm21414@st.uel.edu.vn
5	Nguyễn Thị Thu Uyên	K214142098	uyenntt21414@st.uel.edu.vn
6	Nguyễn Hoàng Vi	K214142099	vinh21414@st.uel.edu.vn

Thành phố Hồ Chí Minh, tháng 5 năm 20223

MỤC LỤC

Tóm tắt đồ án

Cho vay vốn là một hoạt động kinh doanh quan trọng đối với cả cá nhân và các tổ chức tài chính. Lợi nhuận và thiệt hại của người cho vay tài chính ở một mức độ nào đó phụ thuộc vào việc hoàn trả khoản vay. Mặc dù cho vay có lợi cho cả người cho vay và người đi vay, nhưng nó tiềm ẩn rủi ro lớn về việc người vay không có khả năng trả lại khoản vay hay còn được gọi là vỡ nợ. Dự đoán khả năng vỡ nợ của khoản vay là một bước quan trọng nên được thực hiện bởi những người cho vay tài chính để giúp họ tìm hiểu xem một khoản vay có thể được hoàn trả hay không. Dự đoán vỡ nợ chính xác có thể giúp các tổ chức tài chính giảm số lượng các khoản cho vay khó đòi và cuối cùng là tăng lợi nhuận. Mục đích của bài nghiên cứu này là sử dụng các kỹ thuật khai thác dữ liệu để đưa ra dự đoán, phân tích hợp lý về vỡ nợ khoản vay. 5 thuật toán học máy có giám sát được áp dụng để dự đoán khả năng vỡ nợ của khoản vay trong báo cáo này bao gồm: XG-Boost, Gradient Boosting, Random Forest, Logistic Regression và Decision Trees. Nhóm đã có mô hình đạt được độ chính xác cao nhất là 96% khi sử dụng Random Forest.

Các từ khóa : XG-Boost, Gradient Boosting, Random Forest, Logistic Regression, Decision Trees

1. Giới thiệu (Introduction)

Với sức cạnh tranh ngày càng tăng trong sự phát triển vượt bậc của các nền kinh tế và những hạn chế về mặt tài chính, việc vay vốn đã dần trở nên phổ biến hơn. Mặc dù hoạt động cho vay khá có lợi cho cả người cho vay và người vay, được coi là một phần thiết yếu của các giao dịch tài chính, nhưng nó tiềm ẩn một số rủi ro lớn. Rủi ro này được gọi là rủi ro tín dụng hoặc vỡ nợ.

Lãi hay lỗ của bên cho vay tài chính phần lớn phụ thuộc vào việc trả nợ, nghĩa là khách hàng có trả nợ hay không (vỡ nợ). Vì vậy, khi các khoản vay bị vỡ nợ, các tổ chức tài chính sẽ bị thua lỗ, thậm chí có thể dẫn đến phá sản và sụp đổ tổ chức. Bằng cách dự đoán khả năng vỡ nợ của khoản vay, các tổ chức tài chính (người cho vay) có thể giảm rủi ro tín dụng, ngăn ngừa khả năng vỡ nợ của khoản vay và tăng lợi nhuận bằng cách đánh giá khả năng của người đi vay trong việc thực hiện nghĩa vụ trả nợ của họ, tức là dự đoán khả năng vỡ nợ của khoản vay. Quá trình dự báo thời điểm khoản vay sẽ bị vỡ

nợ hay không ban đầu được thực hiện thủ công hoặc bán thủ công. Với sự tiến bộ của tính toán thống kê, một số thuật toán học máy được sử dụng để tính toán và dự đoán khả năng vỡ nợ của khoản vay bằng cách đánh giá dữ liệu lịch sử cho vay của một cá nhân. Trong bài báo cáo này, nhóm giải quyết vấn đề này bằng cách xây dựng mô hình phân loại học máy hiệu suất cao để dự đoán khả năng vỡ nợ của khoản vay.

Nghiên cứu này nhằm chứng minh ứng dụng của phân tích dữ liệu và học máy trong ngành tài chính. Nhóm sử dụng các mô hình học máy để đưa ra dự đoán chính xác về chi tiết khoản vay. Năm thuật toán học máy có giám sát được nhóm áp dụng để dự đoán khả năng vỡ nợ của khoản vay và đạt được độ chính xác cao nhất là 96% khi sử dụng mô hình Random Forest.

Cấu trúc của báo cáo như sau: Phần 2 trình bày các nghiên cứu liên quan; Phần 3 nói về nền tảng lý thuyết; Phần 4 giúp mô tả tập dữ liệu nghiên cứu, tiền xử lý dữ liệu và trích xuất đặc trưng và một số phân tích dữ liệu. Phần 5 là kết quả thử nghiệm và phân tích. Cuối cùng, Phần 6 kết luận bài báo cáo.

2. Các nghiên cứu liên quan (Related work):

Hiện nay các nghiên cứu và đánh giá rủi ro tín tín dụng để cải thiện dự đoán khả năng vay giúp các cơ quan ngân hàng và công ty tài chính chọn được khách hàng đủ điều kiện với mức rủi ro tín dụng thấp. Dự đoán vỡ nợ khoản vay là một chủ đề được nhắc nhiều trong lĩnh vực tài chính ngân hàng. Chủ đề này đã và đang thu hút được được nhiều sự chú ý và quan tâm nghiên cứu hơn.

Trong những năm gần đây, nó đã thu hút nhiều sự tập trung hơn vào nghiên cứu về dự đoán khoản vay và đánh giá rủi ro tín dụng.Do nhu cầu vay ngày càng nhiều nên chúng ta cần cải tiến hơn nữa trong các mô hình chấm điểm tín dụng và dự đoán khả năng vỡ nợ.Các chuyên gia, nhà nghiên cứu đã áp dụng nhiều kỹ thuật về chủ đề này. Nhiều kết luận đáng chú ý đã được rút ra và từ đó là tiền đề cho các nghiên cứu và tìm hiểu.

Thuật toán khai thác được Alomari và Fingerman đưa vào trong nghiên cứu [1] vào năm 2017. Nghiên cứu đưa ra kết luận rằng mô hình phân loại hiệu quả nhất đã đạt được bằng cách sử dụng Random Forest và độ chính xác của nó là 71,75%.

Nghiên cứu [2] của Anand, Velu, Whig và Engineering đã sử dụng thành công nhiều thuật toán Classification để dự đoán vỡ nợ ngân hàng, quá trình phân tích được thực hiện bằng Python và các chỉ số hiệu suất như độ chính xác, khả năng thu hồi, độ chính xác và điểm f1. Nghiên cứu sử dụng mô hình Prediction để phát hiện các khách hàng có vấn đề trong số các khách hàng đi vay.

Aslam, Tariq Aziz, Sohail, Batcha và Nanoscience trong nghiên cứu [3] đã sử dụng Machine Learning để dự đoán rủi ro tín dụng bằng cách chấm điểm tín dụng. Nghiên cứu nhận ra có rất ít nghiên cứu tập trung vào tác động của các tiêu cực sai có thể gây bất lợi cho các công ty cho vay.

Nghiên cứu [4] của Boyapati và Aygun cho thấy cách tiếp cận GVC nhanh hơn gần 40 lần so với các phương pháp Graph - Based Clustering biến hiện có mà còn giữ lại phương sai nhiều hơn 5% so với các gói hiện có. Các dự đoán về bộ tính năng từ GVC chính xác đến 98% khi sử dụng thuật toán XGBoost.

Phương pháp khai thác dữ liệu và thuật toán Machine Learning được đưa vào ở nghiên cứu [5] của Coşer, Maer-matei, Albu, Studies, và Research là một mô hình sử dụng các bộ phân loại như LightGBM, XGBoost, Logistic Regression và Random Forest để đánh giá xác suất vỡ nợ khi tham gia khoản vay của khách hàng. Nghiên cứu thu được kết quả Random Forest được áp dụng cho lấy mẫu dưới mức kịch bản kết hợp, với AUC đại diện là 0,89.

Các tác giả trong nghiên cứu [6] xử lý vấn đề mất cân bằng dữ liệu để nâng cao hiệu suất của dự báo vỡ nợ khoản vay. Phương pháp lấy mẫu kết hợp kết hợp phân cụm, đo độ nhạy ngẫu nhiên và RBF.

Nghiên cứu [7] của tác giả Datkhile, Chandak, Bhandari, Gajare và Karyakarte đã xem xét 4 mô hình học tập được xây dựng và 9 thuộc tính để dự đoán rủi ro tín dụng của tiêu dùng.

Eweoya, Adebiyi, Azeta và Azeta tác giả của nghiên cứu [8] đã sử dụng thuật toán Machine Learning để cho thấy sự bất thường của việc nhận tín dụng và kết thúc trong tình trạng vỡ nợ gây bất lợi cho người cho vay.

Trong nghiên cứu [9] đã sử dụng Random Forest (RF), XGBoost, Adaptive Boosting (AdaBoost), Categorical Boosting (CatBoost) và Light Gradient Boosting Machine (LightGBM) làm thuật toán sử dụng các thuật toán này như phân loại để dự đoán xác suất vỡ nợ của khoản vay là tốt.

Jiang và Systems [10] đưa ra phương pháp Machine Learning để giải quyết vấn đề này. Để nhân rộng học sâu trong nhiều lĩnh vực cần có một kỹ thuật chính quy hiệu quả tên là quy chuẩn hoá nhãn lộn xộn.

Nghiên cứu [11] của các tác giả W. Li, Ding, Chen và Yang đã đưa XGBoost được sử dụng cho học tập đồng bộ và XGBoost mạng nơ-ron sâu và hồi quy logistic sau đó được coi là những người học riêng lẻ không đồng nhất để trải qua quá trình hợp nhất có trọng số tuyến tính.

X. Li et al. và các cộng sự đã sử dụng mô hình hợp nhất dựa trên Hồi quy logistic, Random Forest và CatBoost cho vay dự đoán mặc định [12]. Kết quả dự đoán của mô hình hợp nhất được so sánh với ba mô hình đơn lẻ khác, độ chính xác phân loại và hiệu suất của mô hình hợp nhất là tốt hơn hơn ba mô hình còn lại và nó có thể đạt được dự đoán vỡ nợ một cách hiệu quả và giảm khả năng vỡ nợ tín dụng của khách hàng mà các doanh nghiệp cho vay trực tuyến phải đối mặt.

Thuật toán Machine Learning trong nghiên cứu [13] để xác định chính xác liệu một người, với một số thuộc tính nhất định, có khả năng cao sẽ không trả được khoản vay. Trình phân loại Random Forest cung cấp cho nghiên cứu độ chính xác 80% trong khi phương pháp Decision Tree cung cấp cho nghiên cứu độ chính xác 73%. Do đó, mô hình Random Forest dường như là một lựa chọn tốt hơn cho loại dữ liệu như vậy.

Các tác giả Netzer, Lemaire và Herzenstein [14] đã sử dụng Text - mining và Machine Learning để tự động xử lý và phân tích văn bản thô trong hơn 120.000 yêu cầu cho vay từ Prosper. Bao gồm trong mô hình dự đoán, thông tin văn bản trong khoản vay giúp dự đoán đáng kể khả năng vỡ nợ của khoản vay và có thể có ý nghĩa tài chính đáng kể.

Thuật toán Random forest, logistic regression, decision tree, KNN algorithms và naive-biasalgorithm được đưa vào nghiên cứu [15] .Trong số các công cụ phân loại dựa trên cây, rừng ngẫu nhiên vượt trội hơn cây quyết định với độ chính xác 0,3% và với rừng

ngẫu nhiên có AUC cao nhất, chúng tôi có thể kết luận rằng rừng ngẫu nhiên được sử dụng là mô hình phù hợp nhất cho kịch bản này.

Trong nghiên [16] các tác giả Song, Wang, Ye, Zaretzki và Liu đã phát triển một phương pháp phân loại tập hợp đa mục tiêu và cụ thể theo xếp hạng mới cho nhiệm vụ đánh giá rủi ro tín dụng mất cân bằng. Phương pháp gợi ý bắt đầu từ việc xác định bài toán tối ưu đa mục tiêu. Đối với các hồ sơ cho vay trong một danh mục xếp hạng tín dụng cụ thể.

Stevenson, Mues và Bravo đã đưa phương pháp Deep Learning vào trong nghiên cứu [17] rất hữu ích cho dự đoán mặc định của mSME, vì trong các mẫu thử nghiệm phương pháp Deep Learning mới sử dụng mô hình BERT vượt trội so với hai mô hình chuẩn – Logistic Regression và Random Forests. Kết quả hiệu suất của nghiên cứu cho thấy rằng thông tin cho vay dạng văn bản có khả năng dự đoán, tạo ra các dự đoán khá chính xác.

Trong nghiên cứu [18] nghiên cứu ảnh hưởng của điểm tín dụng Zhima đối với dự đoán vỡ nợ đối với các khoản vay cá nhân. Nghiên cứu thiết kế và thực hiện một thử nghiệm trên tập dữ liệu trong đó có hơn 20 nghìn khoản vay cá nhân trực tuyến và người vay Điểm tín dụng Zhima có liên quan. Kết quả thử nghiệm cho thấy việc sử dụng Điểm tín dụng Zhima thực sự có thể cải thiện hiệu suất phân loại của dự đoán mặc định và mức độ cải thiện là khác nhau.

Trong nghiên cứu [19] các tác giả Wang, Chen và Da đã đề xuất một quan điểm mới cho dự đoán vỡ nợ khoản vay dựa trên lợi nhuận, đó là sử dụng BO để tối ưu hóa các siêu tham số của CBT và mục tiêu tối ưu hóa được thiết lập một cách sáng tạo làm chỉ báo lợi nhuận (tức là APR).

Thuật toán CNN-LightGBM được đề xuất trong nghiên cứu [20] được xác minh bằng cách so sánh nó với một mô hình duy nhất. Kết quả cho thấy so với các thuật toán hồi quy logistic, XGBoost và LightGBM, mô hình CNN-LightGBM có độ chính xác và hiệu quả phân loại cao hơn trong dự báo vỡ nợ khoản vay. Kết quả dự đoán cao hơn 90% và AUC cao hơn 95%, điều này xác minh tính khả thi của phương pháp này.

3. Nền tảng lý thuyết (Background)

Hiện nay hầu như mọi lĩnh vực trên thế giới đều đang tiến tới tự động hóa hoàn toàn. Ngày càng có nhiều các phương pháp khác nhau được phát triển nhằm đạt được mục

tiêu trong từng lĩnh vực nghiên cứu được đề ra qua nhiều năm. Nói đến một trong những lĩnh vực tiên tiến nhất, đã thu hút sự chú ý và hứng thú của các nhà khoa học, nhà nghiên cứu và nhà công nghệ không thể không kể đến Trí tuệ nhân tạo (AI).

AI là ý tưởng tạo ra một máy tính hoặc máy móc để mô phỏng trí thông minh giống con người và các hành vi. Nó bắt nguồn từ thời điểm máy tính được chế tạo lần đầu tiên và kể từ đó đã đa dạng hóa thành các lĩnh vực khác nhau như học máy (Machine Learning), mạng thần kinh (Neural Networking), xử lý ngôn ngữ tự nhiên (NLP),...

3.1. Các khái niệm

3.1.1. Machine Learning là gì?

- Đó là một khái niệm cho phép máy móc học hỏi thông qua các tương tác, quan sát trong thế giới thực và hành xử như con người từ đó cải thiện khả năng học hỏi và thực hiện bằng cách sử dụng dữ liệu được cung cấp làm đầu vào bên trong. Những năm gần đây, Machine Learning đã thu hút được sự tập trung và quan tâm rất lớn của các nhà nghiên cứu và nhà công nghệ. Việc triển khai các mô hình và thuật toán học máy khác nhau trong các lĩnh vực sẽ tạo ra nhiều nhiệm vụ quan trọng và cuộc sống của con người từ đó cũng sẽ trở nên dễ dàng hơn rất nhiều.
- Hai ví dụ phổ biến được nêu ra ở đây là lĩnh vực ngân hàng và tài chính. Với sự trợ giúp của các mô hình Machine Learning khác nhau, các ngân hàng và công ty tài chính đã có thể tiến hành quan sát các mẫu và đưa ra kết luận trong các lĩnh vực như gian lận thẻ tín dụng, dự đoán vỡ nợ… Nhờ có Machine Learning mọi quá trình trở nên chính xác và dễ dàng hơn nhiều.
- Trên thực tế rất khó để có thể tổng hợp hay cung cấp tất cả các phương thức Machine Learning. Một vài mô hình được đề cập dưới đây dựa trên các phương pháp học máy khác nhau. Thông thường, tên của một mô hình là sự kết hợp của cấu trúc dữ liệu, thiết kế, công cụ ước tính, cơ chế tập hợp và hơn thế nữa. Dưới đây, một vài thuật toán trong lĩnh vực học máy đã được sử dụng là Decision Tree, XG Boost và Random Forest.

3.1.2. Cohen's Kappa

- Cohen's Kappa là một thước đo thống kê được sử dụng để đo lường độ tin cậy của các bên đối với các hạng mục phân loại (định tính) với cùng một số lượng và tần suất những người đánh giá đồng ý với nhau. Nó cũng có thể được sử dụng để đánh giá hiệu suất của một mô hình phân loại.
- Công thức để tính toán hệ số kappa là:

  k = (po – pe) / (1 – pe)

  po : Thỏa thuận được quan sát tương đối giữa những người đánh giá.

  pe : Xác suất giả định của thỏa thuận ngẫu nhiên.
- Giá trị của Cohen's Kappa luôn nằm trong khoảng từ 0 đến 1 trong đó:
- 0 cho biết không có thỏa thuận nào giữa hai người đánh giá.
- 1 chỉ ra sự đồng thuận hoàn hảo giữa hai người đánh giá.

3.1.3. K-fold Cross-validation

- Cross-validation (hay xác thực chéo) là một phương pháp thống kê có sai lệch thấp thường được sử dụng trong học máy nhằm so sánh và chọn một mô hình cho một vấn đề lập mô hình dự đoán nhất định.
- Trong quy trình sử dụng mẫu để đánh giá các mô hình máy học trên một mẫu dữ liệu hạn chế có một tham số duy nhất được gọi là k, đề cập đến số lượng nhóm mà một mẫu dữ liệu nhất định sẽ được chia ra trong quy trình. Vì thế, nó được gọi là quy trình xác thực chéo k-fold.

3.2. Các mô hình sử dụng trong dự án

Decision Tree

- Là một thuật toán linh hoạt được sử dụng để thực hiện các nhiệm vụ phân loại và hồi quy. Là một trong số các thuật toán phổ biến nhất được sử dụng để phân loại bao gồm một số nhánh, lá và gốc. Thuật toán tạo ra một cấu trúc có hình dạng giống như một cái cây bằng cách phân loại các thể hiện và sử dụng một thuật toán phân chia đệ quy. Một nhãn lớp sẽ được đại diện bởi một nút lá và các nhánh

đại diện cho kết quả thử nghiệm. Các thử nghiệm này được đại diện bởi các nút bên trong cho một thuộc tính.

XGBoost

- XGBoost thuộc thuật toán học có giám sát. Tương tự Decision Tree, mục đích sử dụng của chúng là phân loại, hồi quy hoặc giải quyết các vấn đề do người dùng tự định nghĩa. Một nhóm dự đoán được xây dựng với một số cây quyết định mở rộng trong không gian con, dữ liệu được chọn ngẫu nhiên và được xây dựng song song. Vấn đề cốt lõi của thuật toán này là tối ưu hóa giá trị của hàm mục tiêu. Nó tuân thủ một chiến lược theo cấp độ và thực hiện các thuật toán học máy theo khung tăng cường độ dốc. Sau đó, sử dụng các tổng từng phần này để đánh giá chất lượng của các phần có thể có trong tập huấn luyện.

Random Forest

- Thuật toán RF được xây dựng trên nền tảng mô hình Decision Tree. Ở đây mỗi cây có nhiệm vụ như một phiếu làm cơ sở ra quyết định. Cùng với các kết quả riêng lẻ, các phương pháp học nhóm cũng được kết hợp để mang lại một kết quả có độ tin cậy cao hơn. Đồng thời, dựa trên kỹ thuật đóng gói (bagging) hoặc tập hợp (boostrap), Random Forest được mở rộng và sử dụng các mẫu ngẫu nhiên (có lặp lại) của dữ liệu huấn luyện nhằm tạo ra nhiều dữ liệu hồi quy không cần chỉnh sửa và là tổng kết quả trung bình của chúng.

Logistic Regression

- Còn được gọi là mô hình logit, mô hình này thường được sử dụng để phân loại và phân tích dự đoán. Hồi quy logistic ước tính xác suất xảy ra một sự kiện, dựa trên tập dữ liệu nhất định gồm các biến độc lập. Vì kết quả là một xác suất nên biến phụ thuộc có giới hạn từ 0 đến 1. Trong hồi quy logistic, phép biến đổi logit được áp dụng theo tỷ lệ, điều đó nghĩa là xác suất thành công chia cho xác suất thất bại. Điều này còn thường được gọi là tỷ lệ cược log hoặc logarit tự nhiên của tỷ lệ cược được biểu thị dưới dạng công thức như sau:

  $\text{Logit}(pi) = 1/(1+ \exp(-pi))$

  $\ln(pi/(1-pi)) = \text{Beta}_0 + \text{Beta}_1*X_1 + \ldots + B_k*K_k$

Gradient Boosting

- Là một kỹ thuật học máy cho các vấn đề hồi quy và phân loại tạo ra một mô hình dự đoán dưới dạng một tập hợp các mô hình dự đoán yếu. Kỹ thuật này xây dựng một mô hình theo kiểu từng giai đoạn và tổng quát hóa mô hình bằng cách cho phép tối ưu hóa một hàm mất mát khả vi tùy ý. Giống như các mô hình khác, Gradient Boosting về cơ bản kết hợp nhiều learner yếu để tạo thành một learner mạnh theo kiểu lặp đi lặp lại. Khi mỗi learner yếu được thêm vào, một mô hình mới được trang bị để cung cấp ước tính chính xác hơn về biến phản hồi. Ý tưởng của Gradient Boosting chính là có thể kết hợp một nhóm các mô hình dự đoán tương đối yếu để xây dựng một mô hình dự đoán mạnh hơn.

4. Phương pháp luận nghiên cứu

4.1. Thu thập dữ liệu

Bộ dữ liệu được lấy từ Kaggle "Loan Defaulter" gồm hai file dữ liệu tên là application_data, previous_application.

- Tập dữ liệu "application_data" gồm thông tin của 307511 khách hàng về việc xin vay với tổng cộng 122 thuộc tính. Các thuộc tính này được sử dụng để mô tả thông tin về khách hàng, bao gồm như độ tuổi, nguồn thu nhập, tình trạng hôn nhân, tài sản sở hữu, mục đích vay, lịch sử tín dụng,.. Bên cạnh đó, tập dữ liệu này còn chứa thuộc tính "TARGET", đây là thuộc tính chính cần được dự đoán và nó cho biết khách hàng có khả năng trả nợ hay không.
- Tập dữ liệu "previous_application" chứa thông tin của 1670214 khách hàng và 37 thuộc tính về quá trình đăng ký vay trước đó. Các thuộc tính này chứa thông tin về số tiền vay, mục đích vay, nơi làm việc, tần suất thanh toán, thời gian xét duyệt đơn vay,.. Thông tin từ tập tin này có thể được sử dụng để tạo mô hình dự đoán khả năng không trả nợ của khách hàng trong tương lai.

4.2. Mô tả dữ liệu

Phần mô tả dữ liệu được lấy từ Kaggle "Loan Defaulter" tên là columns_description chứa phần giải thích ý nghĩa về tất cả các thuộc tính trong tập "application_data" và "previous_application". Nó cung cấp cho người dùng

cái nhìn sâu hơn về dữ liệu và có thể giúp cho việc hiểu và phân tích một cách chính xác hơn.

Table	Variable name	Description	Variable type
application_data	TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)	float64
application_data	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving	object
application_data	CODE_GENDER	Gender of the client	object
application_data	AMT_INCOME_TOTAL	Credit amount of the loan	float64
application_data	AMT_CREDIT	Credit amount of the loan	float64
application_data	AMT_ANNUITY	Loan annuity	float64
application_data	NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,…)	object

application_data	NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)	object
application_data	DAYS_EMPLOYED	How many days before the application the person started current employment	int64
….	….	….	….
previous_application	AMT_DOWN_PAYMENT	Down payment on the previous application	float64
previous_application	AMT_GOODS_PRICE	Goods price of good that client asked for (if applicable) on the previous application	float64
previous_application	NAME_CONTRACT_STATUS	Contract status (approved, canceled, ...) of previous application	object

4.3. Tiền xử lý dữ liệu và trích xuất đặc trưng

- Xử lý dữ liệu bị thiếu
  - Đổi tên các thuộc tính thành chữ viết thường cho việc thực hiện mô hình dễ dàng hơn
  - Liệt kê ra các cột có thông tin bị thiếu và cho thấy có đến 49 thuộc tính có dữ liệu bị thiếu lên đến hơn 40% và cần phải xóa chúng đi để tránh ảnh hưởng đến kết quả

- Loại bỏ những hàng có giá trị null như trong cột 'days_last_phone_change', 'cnt_fam_members'

- Điền vào dữ liệu bị thiếu
  - Dùng phương pháp nhập giá trị trung vị/giá trị xuất hiện nhiều nhất để xử lí các giá trị null trong các cột của tập dữ liệu. Ở trong bài, nhóm có đưa giá trị trung vị của các cột 'amt_annuity', 'amt_goods_price', 'ext_source_2', 'ext_source_3' vào những ô có giá trị null trong cột tương ứng. Hoặc nhập giá trị mod vào 11 cột như 'name_type_suite', 'obs_30_cnt_social_circle','amt_req_credit_bureau_year','def_30_cnt_social_circle','obs_60_cnt_social_circle','def_60_cnt_social_circle','amt_req_credit_bureau_qrt','amt_req_credit_bureau_mon','amt_req_credit_bureau_week','amt_req_credit_bureau_day','amt_req_credit_bureau_hour'. Chúng đều bị thiếu với cùng một đối tượng dữ liệu nên sẽ được xử lý cùng một cách

- Sửa giá trị không hợp lệ và chuẩn hóa giá trị
  - Thay thế các giá trị XNA trong cột "code_gender". Sử dụng các giá trị hiện có của cột giới tính để xác định tỷ lệ giữa các giá trị 'F' và 'M' tương ứng 0,65 và 0,35 (tức là 65% giá trị được thay thế bằng F và 35% giá trị được thay thế bằng M. Phương pháp này giúp giữ nguyên sự phân bố của các giá trị trong cột và cải thiện độ chính xác của dữ liệu được xử lý. Dùng tương tự phương pháp này với cột "organization_type"
  - Chuyển hóa giá trị ngày sang năm của các cột có liên quan như "days_birth", "days_employed"
  - Tất cả những giá trị của cột "days_birth", "days_employed", "days_id_publish","days_last_phone_change",.. đều mang giá trị âm một cách vô lý nên cần chuyển đổi chúng sang giá trị dương

- Xử lý giá trị ngoại lệ

- Có khoảng 18% giá trị của cột “year_employed” bằng 100. Điều này vô lý bởi vì không có ai có số năm đi làm được 100 cả vì vậy những giá trị này sẽ được thay thế bằng NAN để không ảnh hưởng đến việc phân tích

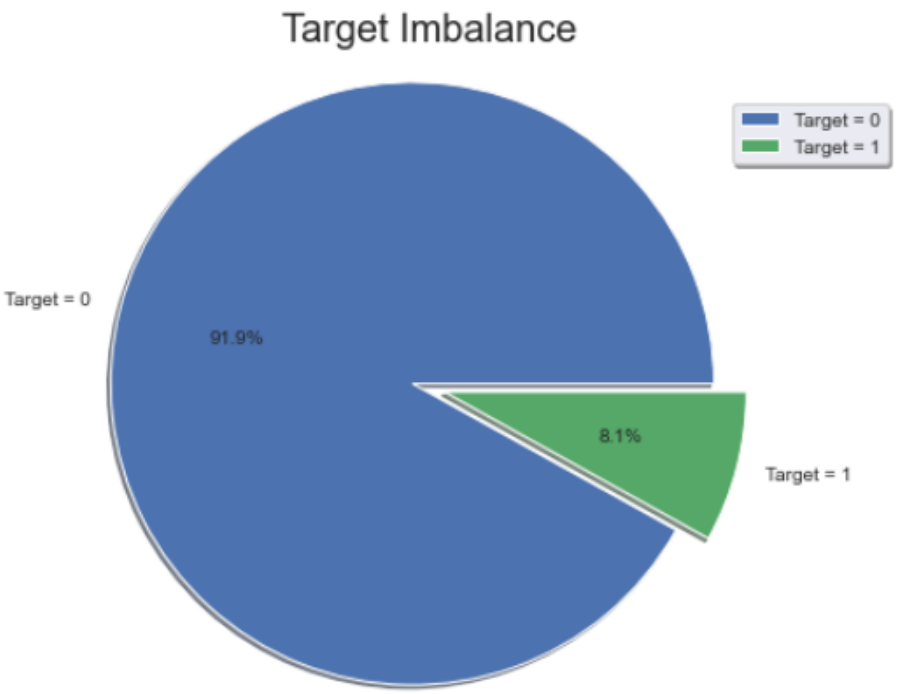
- Mã hóa giá trị bằng phương pháp dùng biến giả
  - Lọc ra các biến định tính và chuyển đổi các biến phân loại thành các giá trị nhị phân (0 hoặc 1) để giúp cho việc chạy mô hình trở nên dễ dàng hơn
- Nhóm đã sử dụng một số kỹ thuật feature engineering như sau:
  - Xóa các giá trị bị thiếu hoặc cập nhật bằng giá trị trung vị, giá trị xuất hiện nhiều nhất để điền vào giá trị bị thiếu
  - Bỏ hơn 40 thuộc tính không cần thiết cho mục đích dự đoán của đề tài như 'days_last_phone_change', 'cnt_fam_members',..
  - Lọc những biến thuộc dữ liệu số ['cnt_children', ‘amt_income_total’, ‘amt_credit_x’, ‘amt_annuity_x’, ‘days_birth’, ‘ext_source_2’, ‘ext_source_3’,‘cnt_fam_members’,‘days_employed’,‘days_id_publish’, ‘cnt_payment’,...] và những biến thuộc dữ liệu phân loại [‘name_contract_type’,‘code_gender’,‘flag_own_car’,‘name_type_suite’ , ‘income_group’, ‘channel_type’,...]
  - Sử dụng phương pháp One-hot encoding để mã hóa và đưa dữ liệu phân loại về dạng số ví dụ như code_gender_M, flag_own_car_Y, flag_own_realty_Y, name_type_suit_x_family,…
- Sử dụng phương pháp Feature Scaling:
  - Thực hiện việc chuẩn hóa dữ liệu trên các cột số của tập dữ liệu bằng cách sử dụng ‘StandardScaler’ để đưa các đặc trưng đầu vào về cùng một phạm vi giá trị. Dùng phương thức ‘fit_transform()’ được gọi trên dữ liệu huấn luyện ‘x_train[num_cols]’ để tính toán giá trị trung bình và độ lệch chuẩn của các đặc trưng số, sau đó chuẩn hóa dữ liệu bằng cách sử dụng các thông số được tính được. Dữ liệu kiểm tra ‘x_test[num_cols]’ sau đó được chuyển đổi bằng cách sử dụng phương thức ‘transform()’ với bộ chuẩn hóa đã được huấn luyện trên dữ liệu huấn luyện. Điều này đảm bảo rằng

quá trình chuẩn hóa dữ liệu là nhất quán giữa dữ liệu huấn luyện và kiểm tra.

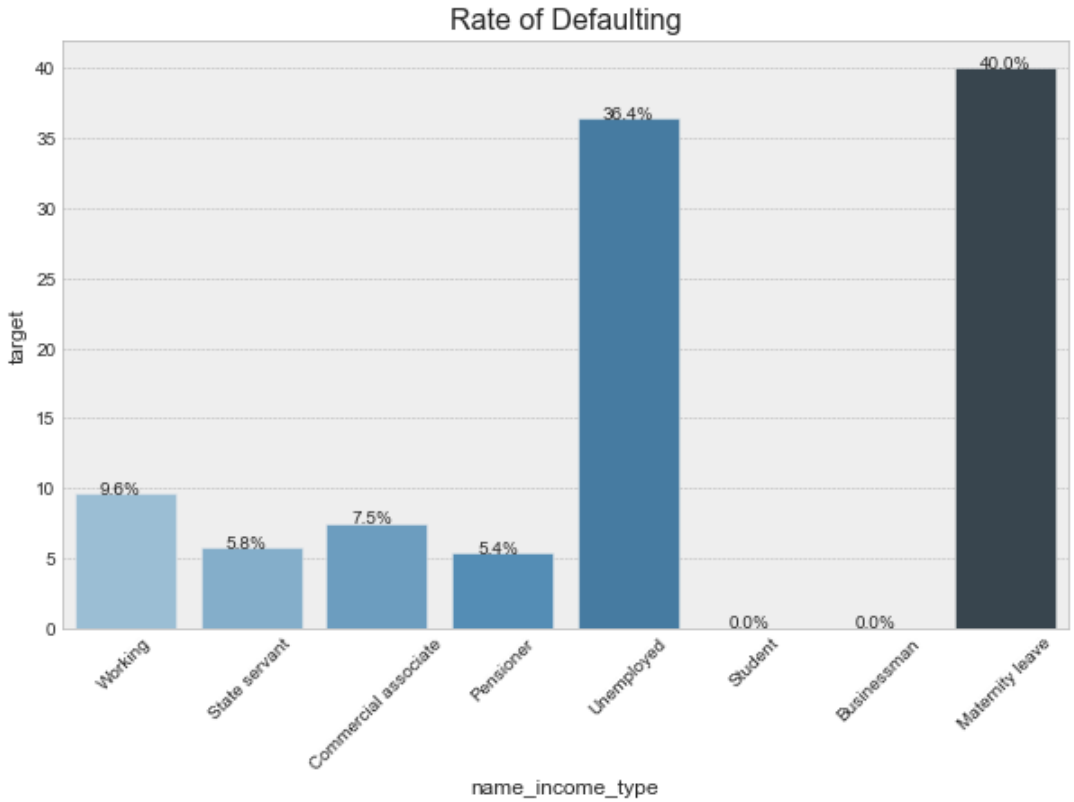
- Class Imbalance: sử dụng SMOTE để cân bằng số lượng mẫu trong mỗi lớp trước khi đi đến phần chạy mô hình và sau khi cân bằng thì cho ra kết quả của hai lớp 0 và 1 đều bằng nhau

-> Tất cả những kỹ thuật này được sử dụng để cải thiện độ chính xác của mô hình và giúp cho mô hình dự đoán khách hàng vay tiền có khả năng trả nợ hay không.

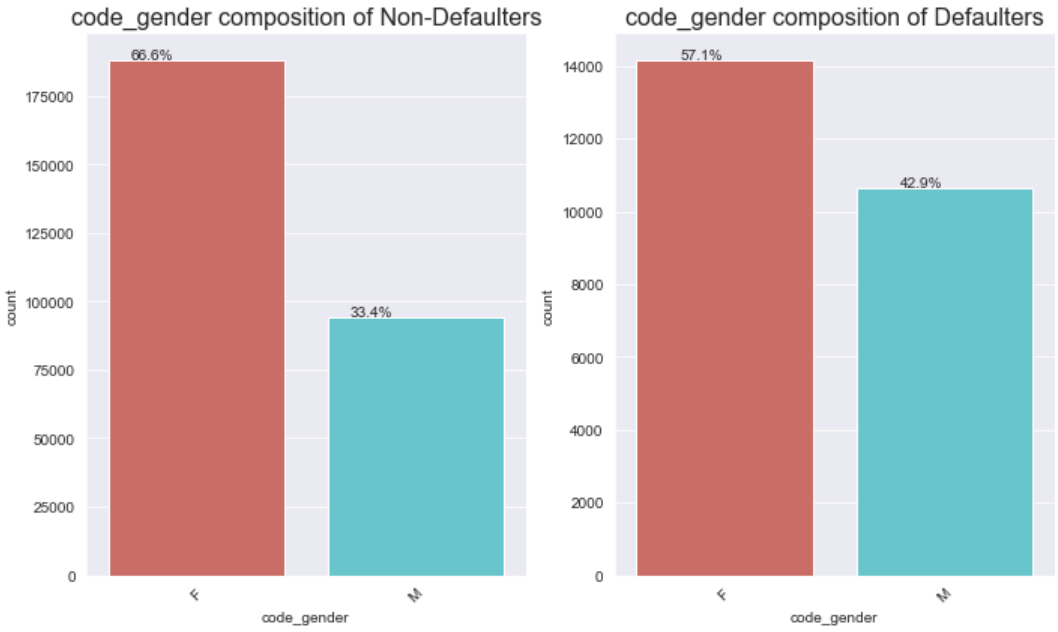
4.4. Một số phân tích



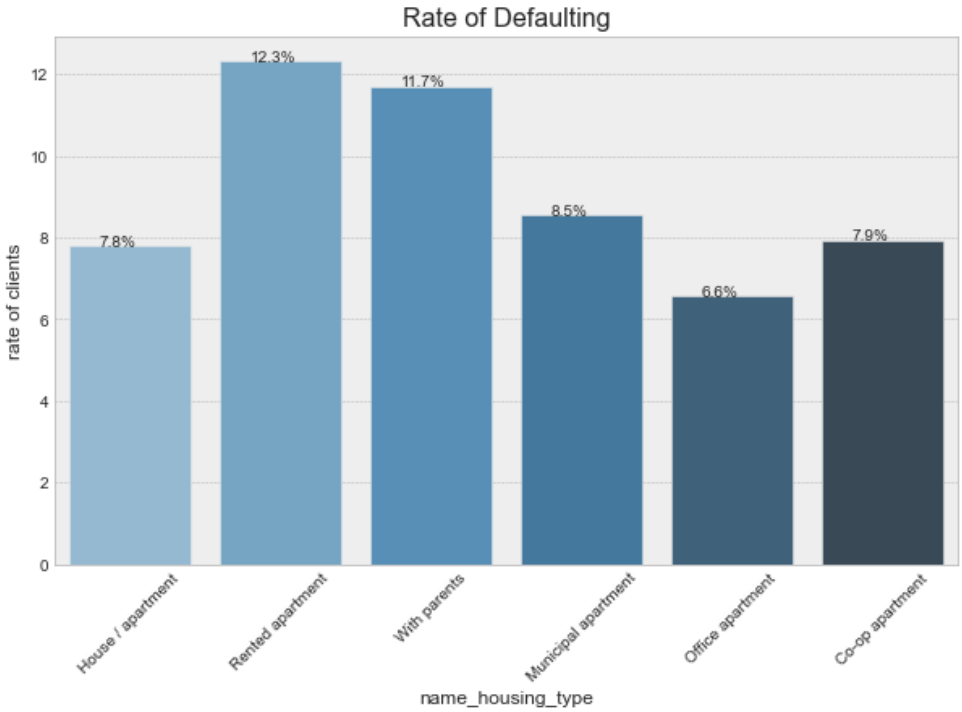
Qua biểu đồ tròn về biến "Target" trên, ta thấy được sự mất cân bằng với 91,9% khách hàng không có lịch sử mặc định và 8,1% khách hàng có lịch sử mặc định. Lịch sử mặc định ở đây tức là tình trạng khách hàng không hoàn trả nợ vay theo đúng thỏa thuận trong hợp đồng. Việc lịch sử mặc định có liên quan đến lịch sử tín dụng, nếu khách đã bị trả nợ chậm hoặc không hoàn trả nợ sẽ để lại lịch sử mặc định. Đây là một yếu tố quan trọng vì nó cho thấy khả năng khách hàng sẽ hoàn trả nợ thế nào trong tương lai nếu có vay tiếp.



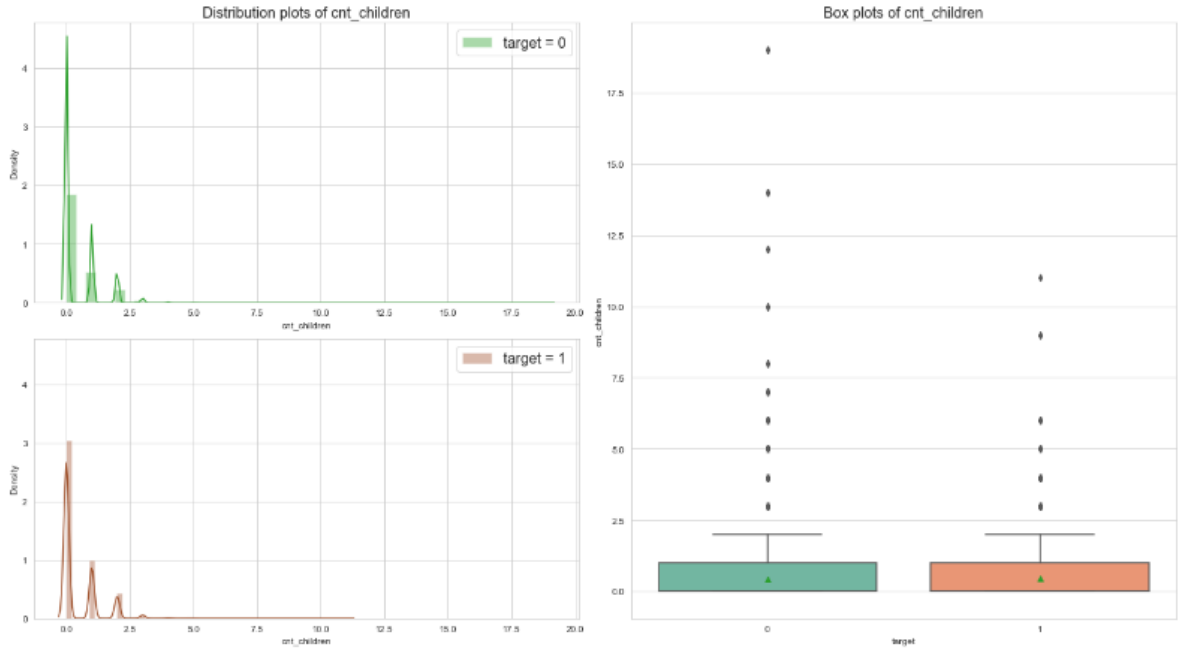
Đây chính là biểu đồ cột kết hợp thể hiện tỉ lệ khách hàng có khó khăn trong việc thanh toán nợ dựa theo từng loại nguồn thu nhập. Tỉ lệ phần trăm tương đối rất ít là những nhóm người thuộc tầng lớp cao, là những người có hưu trí, doanh nhân, khách hàng có công việc ổn định. Hai nhóm khách hàng chiếm tỉ lệ cao nhất lần lượt là nhóm khách hàng thất nghiệp (30,4%) và nhóm khách hàng nghỉ thai sản (40%), cũng dễ hiểu bởi vì họ là những người không có việc làm nên có khả năng để lại nợ xấu là điều không có gì đáng bất ngờ. Tuy nhiên cũng không thể đưa ra sự thuyết phục trong việc dự đoán hai nhóm khách hàng này bởi có rất ít giá trị xuất hiện tương ứng thuộc hai nhóm này.



Biểu đồ cột kết hợp thể hiện tỉ lệ phần trăm chênh lệch giới tính thuộc 2 nhóm khách hàng khác nhau. Nhìn chung ở cả 2 nhóm khách hàng thì tỉ lệ nữ chiếm phần trăm nhiều hơn nam cho thấy nữ nộp đơn xin vay nhiều hơn nam. Tuy nhiên về tình trạng vỡ nợ thì phụ nữ lại ít có khả năng vỡ nợ hơn nam giới, cụ thể là ở nhóm Non-defaulters thì tỉ lệ nữ đang là 66,6% và giảm xuống còn 57,1% khi qua nhóm khách hàng vỡ nợ defaulter và ngược lại với nam giới thì trong nhóm khách hàng vỡ nợ thì lại có tỉ lệ phần trăm tăng lên đến gần 43%.



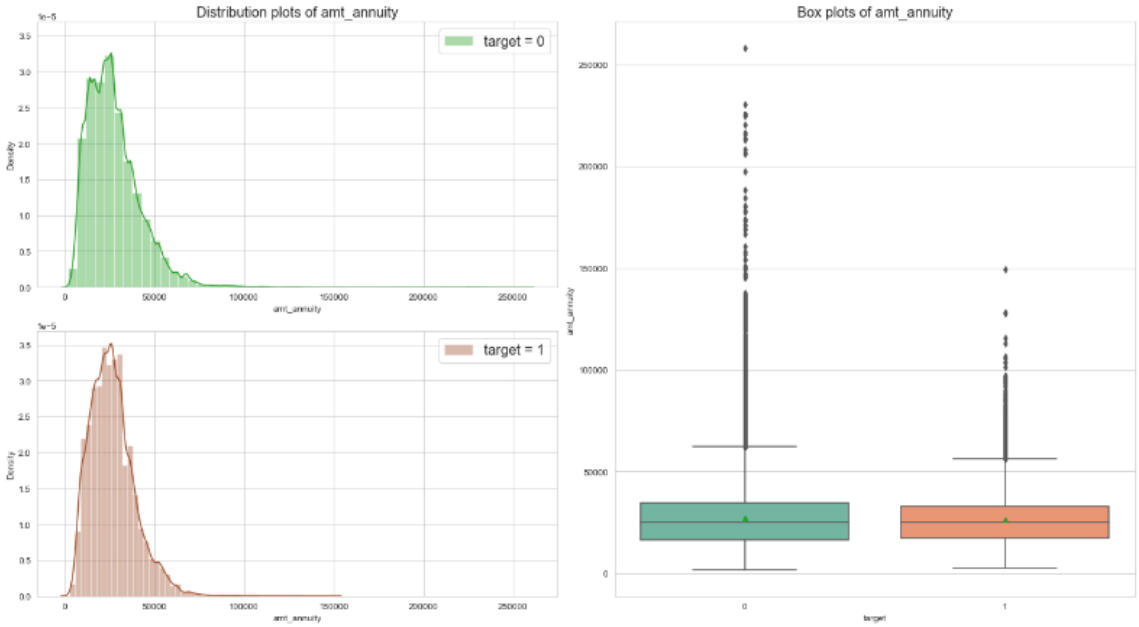
Biểu đồ cột thể hiện tỷ lệ phần trăm khả năng vỡ nợ cho thấy thường là nhóm người chưa có nhà riêng hoặc đang ở với gia đình sẽ chiếm phần trăm cao hơn (trên 11%). Có thể bởi vì họ chưa có đủ tiền để sở hữu cho mình một căn hộ riêng nên họ vẫn còn ở cùng với gia đình hoặc những người đang thuê nhà cũng một phần do chi phí tài chính chưa đủ để chi trả hoặc do những yếu tố khách quan nào đấy ảnh hưởng mà họ cần phải đi vay tiền nhiều hơn. So với nhóm người đã có nhà ở, nhóm văn phòng và chỗ ở kết hợp, căn hộ hợp tác xã, căn hộ cộng đồng thì chi phí sẽ giảm đi phần nào nên lượng người thuộc những nhóm này có tỉ lệ vỡ nợ ít hơn.



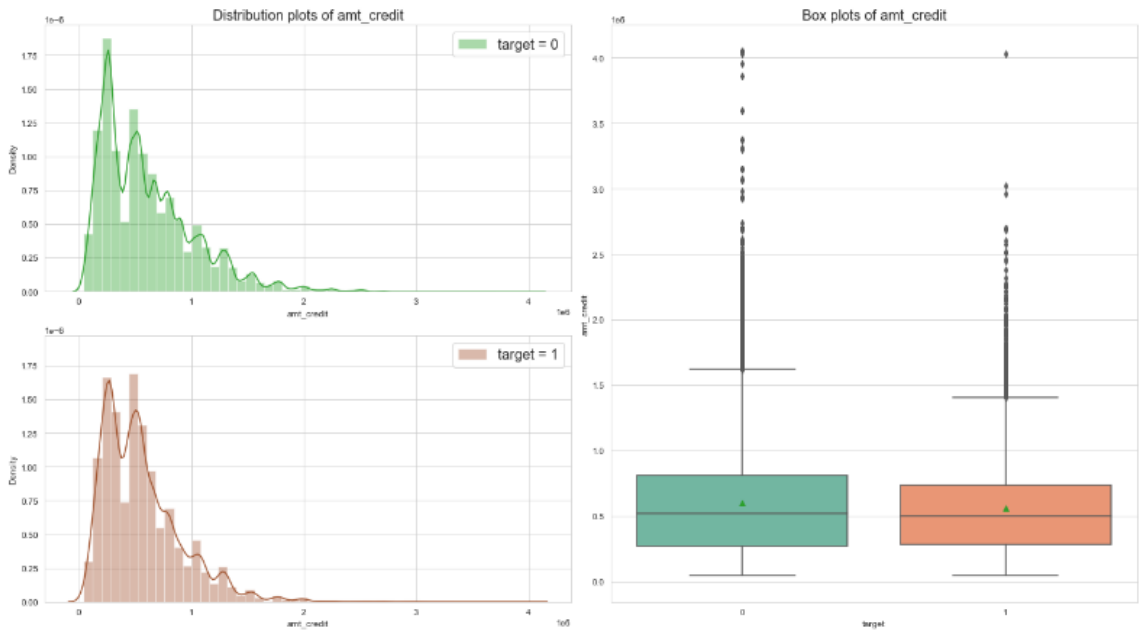
Biểu đồ thể hiện phân bố số lượng con của khách hàng trong tập dữ liệu. Biểu đồ có thể giúp cho việc hiểu và giải thích sự ảnh hưởng của số lượng con đến khả năng thanh toán nợ của khách hàng

Đối với nhóm khách hàng không vỡ nợ (target=0) thì mật độ khách hàng không có con hoặc có tối đa 2 con chiếm đa số và mật độ khách hàng giảm dần khi số lượng con của khách tăng lên, điều này có thể thấy được rõ qua đường biểu diễn trên biểu đồ. Ví dụ trong khoảng giá trị từ 2.5 đến 5.0 ở trục hoành, mật độ khách hàng tương ứng trên trục tung đã giảm rất nhiều so với khoảng giá trị 0 đến 2. Trong khi đó thì nhóm khách hàng vỡ nợ (target=1) có xu hướng có nhiều khách hàng có số lượng trẻ em lớn hơn

Điều này cho thấy rằng số lượng con là một yếu tố ảnh hưởng đến khả năng khách hàng trả nợ đúng hạn. Tuy nhiên, số lượng con không phải yếu tố duy nhất ảnh hưởng đến việc trả nợ mà vẫn còn phụ thuộc bởi nhiều yếu tố khác.



Biểu đồ thể hiện sự phân bố của số tiền trả lãi (amt_annuity) và mật độ khách hàng trong hai nhóm khách hàng. Biểu đồ cho thấy nhóm khách hàng không có khả năng trả nợ thì chỉ số trả lãi tập trung dày đặc ở khoảng giá trị thấp hơn với mật độ nhiều người hơn (0 đến khoảng 30000), có thể hiểu họ thường sẽ vay với mức lãi suất thấp (thời gian vay ngắn) hoặc vay với số tiền cao nhưng thời hạn trả được kéo dài,.. tuy nhiên dù khoản trả lãi hàng tháng có thấp nhưng cũng có thể gây ra rủi ro vỡ nợ cao bởi nếu thời gian vay ngắn là do nhu cầu tạm thời mà không có kế hoạch phân bổ tài chính hiệu quả thì sẽ dễ rơi vào tình trạng nợ chồng chất khó khăn trong việc trả lãi đúng hạn



Biểu đồ thể hiện sự phân bố của số tiền đi vay (amt_credit) và mật độ khách hàng trong hai nhóm khách hàng. Qua biểu đồ, có thể nhận thấy nhóm khách hàng không vỡ nợ (target=0) có phân bố số tiền đi vay tập trung hơn so với nhóm khách hàng vỡ nợ (target=1). Điều này cho thấy khách hàng trong nhóm không vỡ nợ thường có thói quen vay mức tiền nhỏ hơn so với nhóm còn lại. Ngoài ra, nhóm khách hàng vỡ nợ có độ dài đuôi phân bố dài hơn và giá trị trung vị tương đối cao cho thấy dù nhóm này có số lượng khách hàng ít hơn nhưng các khách hàng trong nhóm này lại thường vay số tiền lớn hơn. Điều này có thể suy ra rằng số tiền đi vay là một yếu tố không kém phần quan trọng ảnh hưởng đến khả năng trả nợ đúng hạn

5. Kết quả thử nghiệm và phân tích

Trong bài báo cáo này, chúng em đã sử dụng 5 thuật toán học máy gồm XG-Boost, Gradient Boosting, Random Forest, Logistic Regression và Decision Trees để xây dựng mô hình dự đoán khoản vay và đánh giá rủi ro tín dụng.

Kết quả của tất cả các mô hình được hiển thị bên dưới với báo cáo phân loại và ma trận nhầm lẫn để hiểu rõ hơn về độ chính xác và các điểm số khác của từng mô hình.

5.1 XG-Boost

Mô hình XGBoost cho độ chính xác là 95%

Bảng 1. Classification Report for XG-Boost.

	Precision	***Recall***	***F1 score***	***Support***
0	0.92	0.99	0.95	9103
1	0.99	0.91	0.95	9103
Accuracy			0.95	18247
Macro Avg	0.96	0.95	0.95	18247
Weighted Avg	0.96	0.95	0.95	18247

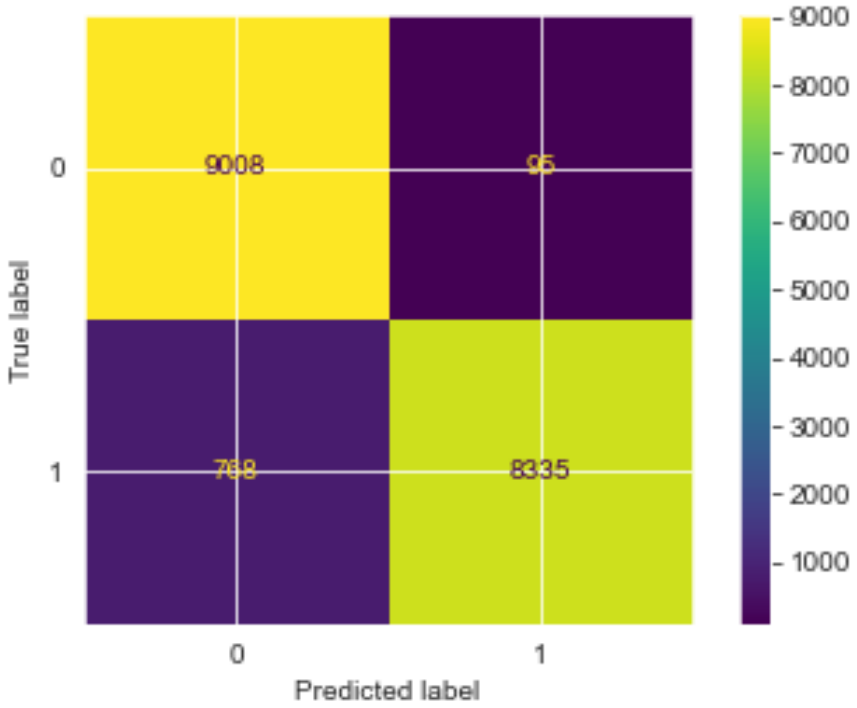


Figure 1. Confusion Matrix of XG-Boost.

5.2 GradientBoosting

Mô hình Gradient Boosting cho độ chính xác là 95%

Bảng 2 . Classification Report for GradientBoosting.

	Precision	***Recall***	***F1 score***	***Support***
0	0.91	1.00	0.95	9103
1	1.00	0.90	0.95	9103
Accuracy			0.95	18206
Macro Avg	0.95	0.95	0.95	18206
Weighted Avg	0.95	0.95	0.95	18206

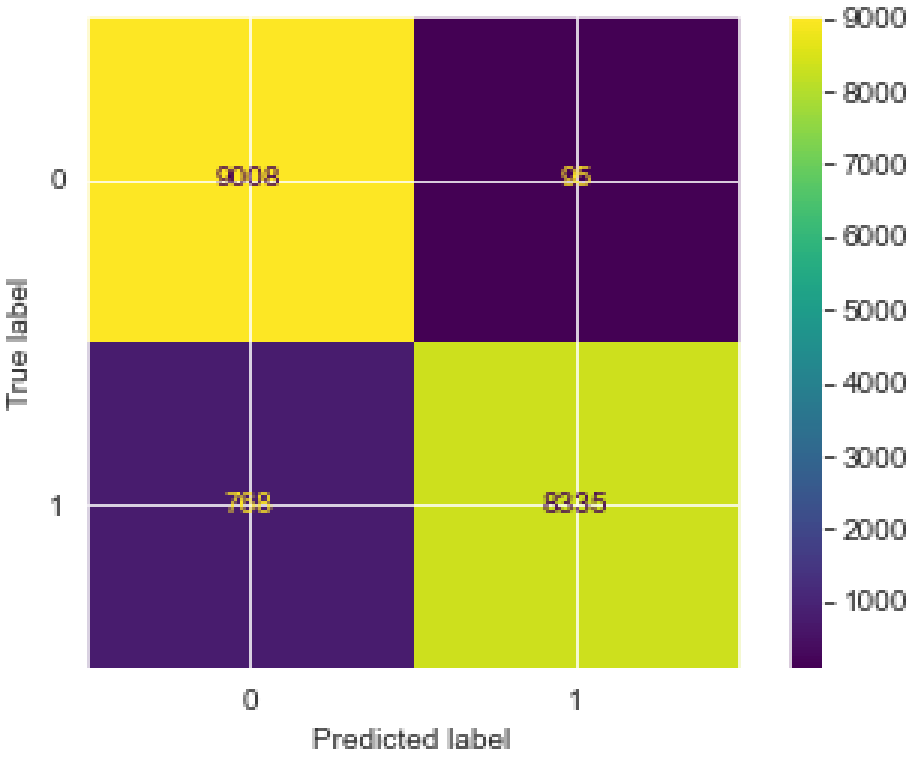


Figure 2. Confusion Matrix of GradientBoosting.

5.3 LogisticRegression

Mô hình Logistic Regression cho độ chính xác là 71%

Bảng 3. Classification Report for LogisticRegression.

	Precision	***Recall***	***F1 score***	***Support***
0	0.72	0.70	0.71	9103
1	0.71	0.72	0.72	9103
Accuracy			0.71	18206
Macro Avg	0.71	0.71	0.71	18206
Weighted Avg	0.71	0.71	0.71	18206

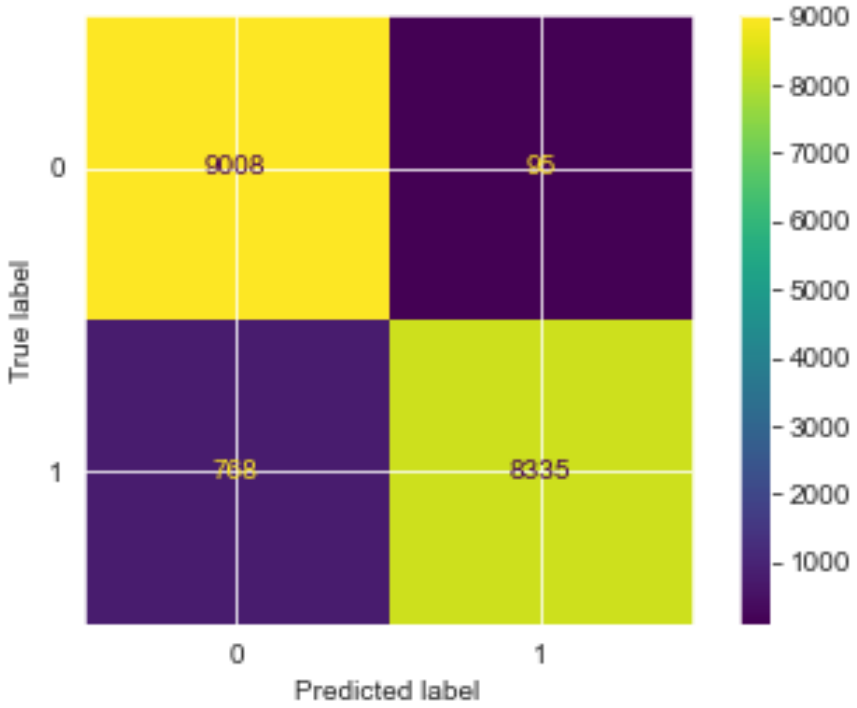


Figure 3. Confusion Matrix ofLogisticRegression.

5.4 Decision Tree

Mô hình Decision Tree cho độ chính xác là 91%.

Bảng 4. Classification Report for Decision Tree.

	Precision	***Recall***	***F1 score***	***Support***
0	0.92	0.89	0.90	9103
1	0.90	0.92	0.91	9103
Accuracy			0.91	18206
Macro Avg	0.91	0.91	0.91	18206
Weighted Avg	0.91	0.91	0.91	18206

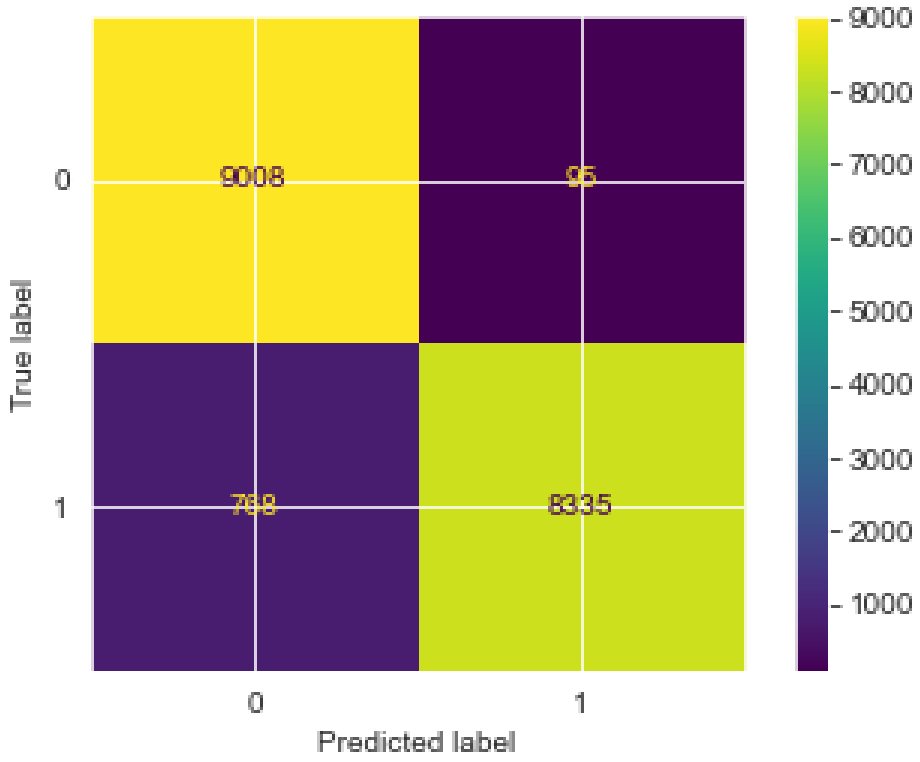


Figure 4. Confusion Matrix of DecisionTree.

5.5 Random Forest

Mô hình Random Forest cho độ chính xác là 96%.

Bảng 5. Classification Report for Random Forest .

	Precision	***Recall***	***F1 score***	***Support***
0	0.92	1.00	0.96	9123
1	1.00	0.92	0.96	9124
Accuracy			0.96	18247
Macro Avg	0.96	0.96	0.96	18247
Weighted Avg	0.96	0.96	0.96	18247

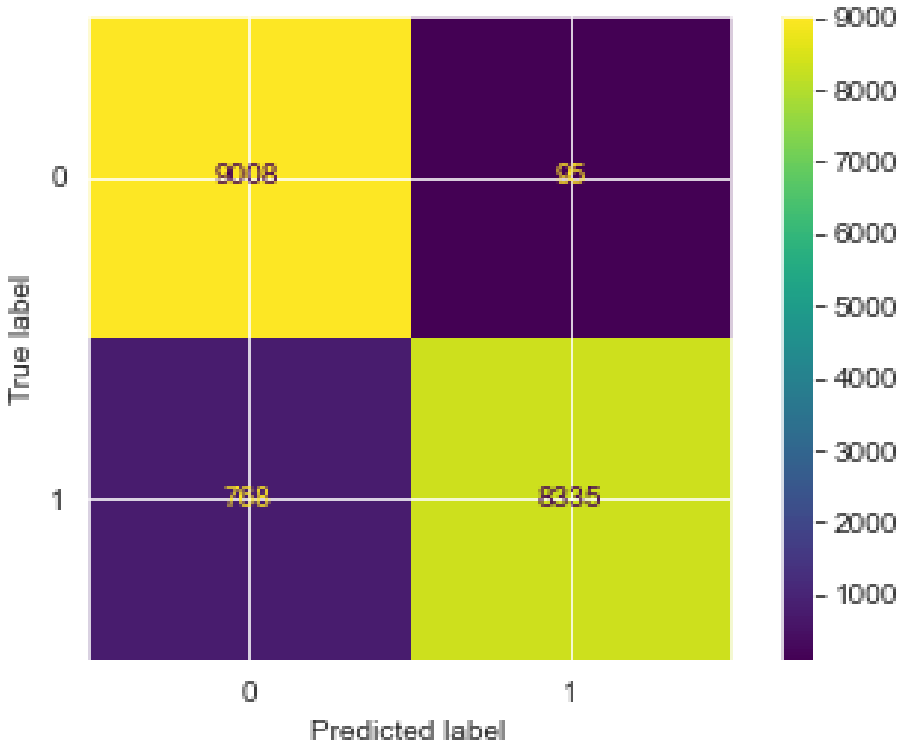


Figure 5. Confusion Matrix of Random Forest.

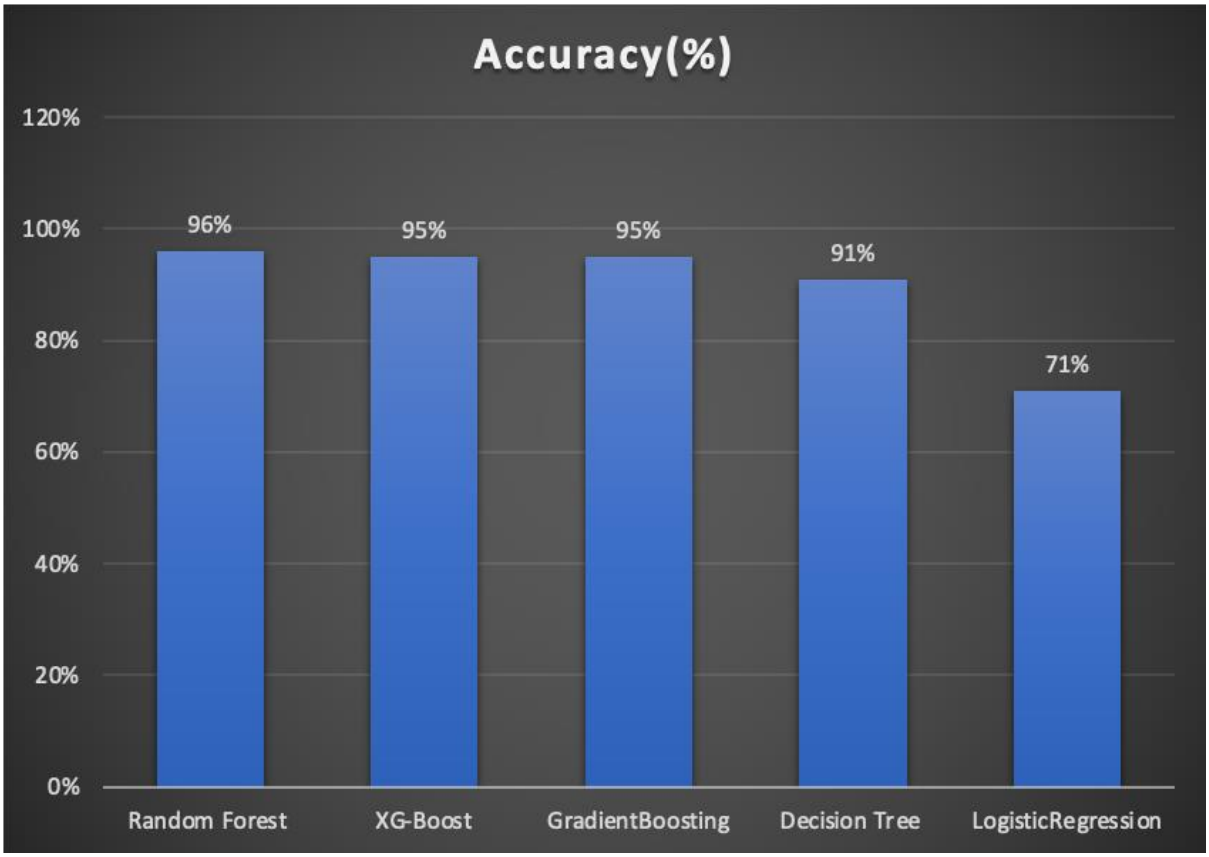
Bảng 6. Hiệu suất trung bình của các mô hình học máy

Model	Accuracy	Cohen's Kappa	Class	Precision	Recall	F1-score
XGBoost	0.95	0.91	**0**	0.92	0.99	0.95
			1	0.99	0.92	0.95
Gradient Boosting	0.95	0.90	**0**	0.91	1.00	0.95
			1	1.00	0.90	0.95
Logistic Regression	0.71	0.42	**0**	0.72	0.70	0.71
			1	0.71	0.72	0.72
Decision Tree	0.91	0.81	**0**	0.92	0.89	0.90
			1	0.90	0.92	0.91
Random Forest	0.96	0.92	**0**	0.93	1.00	0.96
			1	1.00	0.92	0.96

Bảng 6 cho thấy hiệu suất trung bình của các loại máy học như XGBoost, Gradient Boosting, Logistic Regression, Decision Tree và Random Forest. Để quan sát hiệu suất của mô hình, nhóm đã đưa ra kết quả về Accuracy, Cohen's Kappa, Precision, Recall và F1-Score.

Hình 2 cho thấy mô hình Random Forest có điểm chính xác cao nhất là 96 %, trong khi mô hình Logistic Regression có độ chính xác thấp nhất đạt 71%. Ngoài ra, các mô hình như XgBoost, Gradient Boosting và Decision Tree hoạt động khá tốt với số độ chính xác lần lượt là là 95%, 95%, 91%.

Hình 2



Nhìn vào các ma trận nhầm lẫn và báo cáo phân loại của tất cả các mô hình trên, có thể nói rằng thuật toán XG-Boost là một lựa chọn tốt hơn so với các mô hình khác để dự đoán khoản vay trên tập dữ liệu đã cho.

6. Kết luận

Dự đoán các trường hợp vỡ nợ trong thị trường tín dụng cụ thể ở đề tài nghiên cứu của nhóm liên quan đến nền tảng cho vay là một nhiệm vụ quan trọng và đầy thách thức. Qua bài nghiên cứu của nhóm có thể đưa ra một số ưu, nhược điểm cũng như hướng phát triển trong tương lai như sau:

6.1. Ưu điểm

Các mô hình dự đoán chính xác (như ở bài phân tích của nhóm mô hình XGBoost là tốt nhất trong các mô hình) sẽ rất có lợi vì nếu sử dụng mô hình dự đoán không chính xác sẽ dẫn đến sự thất bại của nền tảng cho vay từ đó gây ra một loạt rủi ro tài chính. Kết quả thực nghiệm của nhóm xác nhận rằng giới tính, trình độ học vấn, hôn nhân, thu nhập, công việc và các thông tin xác minh khác đóng vai trò quan trọng trong việc dự đoán khả năng vỡ nợ khoản vay của người đi vay. Hơn nữa, phát hiện của nhóm cũng cho thấy rằng các phương pháp học máy có triển vọng ứng dụng rộng rãi trong dự đoán vỡ nợ cho vay đồng thời cũng cung cấp các kỹ thuật quan trọng cho các cơ quan quản lý và công ty cho vay về sàng lọc người đi vay và quản lý nền tảng, điều này cuối cùng có thể làm giảm rủi ro trực tiếp là làm giảm rủi ro tài chính và thị trường trên toàn thế giới. Kết luận nghiên cứu của nhóm dựa trên dự đoán vỡ nợ trên nền tảng bằng phương pháp học máy - phương pháp được sử dụng rộng rãi trong đánh giá rủi ro tín dụng của người vay để giúp ngân hàng và người vay chống lại rủi ro tài chính một cách hiệu quả. Đặc biệt trong bối cảnh dịch bệnh hoặc những chủng bệnh mới có khả năng bùng phát bất cứ lúc nào ở toàn cầu hiện nay tương tự dịch COVID 19, suy thoái kinh tế toàn cầu là điều hiển nhiên, rủi ro tài chính đang gia tăng và thị trường tài chính bị tác động nghiêm trọng càng làm nổi bật tính hiệu quả và chính xác của các phương pháp học máy. Theo báo cáo năm 2019 của McKinsey, học máy có thể giảm 10% tổn thất tín dụng của ngân hàng và giảm thời gian ra quyết định tín dụng từ 25% đến 50%. Zestfinance - một trong những công ty đầu tiên áp dụng các phương pháp học máy để đánh giá rủi ro tín dụng, nhận thấy rằng hiệu suất của các mô hình học máy cao hơn 40% so với các mô hình đánh giá tín dụng truyền thống. Vì vậy, sâu xa hơn phương pháp nghiên cứu và kết luận của nhóm có thể áp dụng cho các lĩnh vực sau:

- Đầu tiên, các phương pháp và kết luận của nhóm có thể giúp các ngân hàng thiết lập mô hình chấm điểm tín dụng đối với doanh nghiệp vay vốn. Cách tiếp cận của nhóm hiện có thể giúp dự đoán điểm tín dụng và khả năng vỡ nợ của bên vay kết hợp với tình trạng tài sản của bên vay, hồ sơ tín dụng, khả năng sinh lời của hoạt động kinh doanh hiện tại, tốc độ tăng trưởng kinh doanh, vòng quay vốn lưu động, triển vọng phát triển ngành, tình hình hoạt động của các doanh nghiệp trực

thuộc và dữ liệu kiểm toán công thương nghiệp và cục thuế. Trước tiên, chúng ta có thể sử dụng chương trình học máy để phân tích và tiền xử lý sơ bộ dữ liệu liên quan của doanh nghiệp đi vay, sau đó sử dụng phương pháp Random Forest hoặc thuật toán XGBoost để thực hiện lựa chọn tính năng, phân tách dữ liệu, xây dựng mô hình đào tạo học máy và sử dụng mô hình học máy để hoàn thành dự đoán dữ liệu. Ở giai đoạn cuối, để đo lường hiệu quả của mô hình, chúng ta có thể sử dụng các chỉ số đánh giá khác nhau, chẳng hạn như độ chính xác *(precision, recall, f1 score),* cohen-kappa, độ chính xác và khả năng thu hồi *(accuracy)*, để đánh giá hiệu quả dự đoán. Ngoài ra, chúng ta cũng có thể sử dụng các phương pháp học máy để đánh giá tình trạng tín dụng của những người vay cá nhân. Kết hợp với dữ liệu về tài sản và hồ sơ tín dụng của từng người vay, phương tiện truyền thông xã hội như tham gia Twitter hoặc Facebook, thanh toán mạng, nền tảng cá nhân, văn hóa và tôn giáo, mạng xã hội và các đặc điểm nhân khẩu học như giới tính, tuổi tác, nghề nghiệp, giáo dục, có khả năng cung cấp dữ liệu sẽ được sử dụng để dự đoán điểm tín dụng và khả năng vỡ nợ của người đi vay.

- Thứ hai, phương pháp nghiên cứu và kết luận của nhóm cũng mang lại giá trị tham khảo tốt cho các cơ quan quản lý tài chính. Đầu tiên, các cơ quan quản lý có thể sử dụng dữ liệu lớn của doanh nghiệp để dự đoán rủi ro tài chính khu vực và sau đó thực hiện các biện pháp điều tiết, chẳng hạn như giảm hoặc tăng vốn lưu động, tăng hoặc giảm tỷ lệ dự trữ ngân hàng và tăng hoặc giảm lãi suất cho vay của ngân hàng, để đạt được mục đích quản lý rủi ro tài chính.
- Thứ ba, các cơ quan quản lý có thể sử dụng các phương pháp học máy để xác định gian lận tài chính và đưa ra dữ liệu giám sát rủi ro cho người cho vay thông qua các cảnh báo rủi ro. Cụ thể, các cơ quan quản lý có thể trích xuất dữ liệu giao dịch của thẻ tín dụng, sử dụng các phương pháp học máy để đào tạo và kiểm tra ngược, trích xuất các đặc điểm chính của gian lận tài chính và phân biệt các đặc điểm này với các giao dịch thông thường. Cơ chế nhận dạng này giúp cơ quan quản lý xác định hiệu quả hành vi gian lận tài chính, đặc biệt là những hành vi có điều kiện kinh doanh kém lợi dụng việc phòng ngừa và kiểm soát dịch bệnh *(nếu có)* để lừa đảo các khoản trợ cấp của chính phủ.

Ngoài ra, theo kết luận nghiên cứu đề tài, nhóm tin rằng chính phủ có thể quan tâm nhiều hơn đến khách hàng tín dụng cá nhân theo những cách sau để giảm thiểu thảm họa tài chính và phục hồi từ chúng:

- Thứ nhất, xác định nhu cầu vay và khả năng trả nợ của bên vay. Đối với người vay cá nhân có nhu cầu vay nhưng không vay được tiền kịp thời, cần đánh giá lại triển vọng phát triển nghề nghiệp và mạng lưới quan hệ xã hội của người vay bằng phương pháp học máy và quan tâm nhiều hơn đến tương lai hơn là hồ sơ tín dụng, tài sản, tài sản thế chấp và các thông tin quá khứ khác. Đối với những người vay có triển vọng phát triển tốt, chính phủ có thể ban hành các chính sách trợ cấp để giúp người vay vượt qua khó khăn.
- Thứ hai, các đề xuất đầu tư nên được đưa ra cho từng người vay theo mức độ chấp nhận rủi ro và khả năng trả nợ của những người vay khác nhau. Các kênh đầu tư như thị trường chứng khoán, thị trường trái phiếu, ngân hàng và tài chính internet có các dạng rủi ro, điều khoản, chi phí và lợi nhuận khác nhau.
- Thứ ba, các tổ chức tài chính được khuyến khích cấp các khoản vay cho những người vay cá nhân có xếp hạng tín dụng cao dưới hình thức trợ cấp và chính sách ưu đãi cho các tổ chức tài chính

Nhìn chung, các phương pháp nghiên cứu và kết luận của nhóm có thể được sử dụng để tham khảo bởi các ngân hàng, cơ quan quản lý tài chính, chính phủ, người đi vay doanh nghiệp và người vay cá nhân.

Các mô hình dự đoán chính xác (như ở bài phân tích của nhóm mô hình XGBoost là tốt nhất trong các mô hình) sẽ rất có lợi vì nếu sử dụng mô hình dự đoán không chính xác sẽ dẫn đến sự thất bại của nền tảng cho vay từ đó gây ra một loạt rủi ro tài chính. Kết quả thực nghiệm của nhóm xác nhận rằng giới tính, trình độ học vấn, hôn nhân, thu nhập, công việc và các thông tin xác minh khác đóng vai trò quan trọng trong việc dự đoán khả năng vỡ nợ khoản vay của người đi vay. Hơn nữa, phát hiện của nhóm cũng cho thấy rằng các phương pháp học máy có triển vọng ứng dụng rộng rãi trong dự đoán vỡ nợ cho vay đồng thời cũng cung cấp các kỹ thuật quan trọng cho các cơ quan quản

lý và công ty cho vay về sàng lọc người đi vay và quản lý nền tảng, điều này cuối cùng có thể làm giảm rủi ro trực tiếp là làm giảm rủi ro tài chính và thị trường trên toàn thế giới. Kết luận nghiên cứu của nhóm dựa trên dự đoán vỡ nợ trên nền tảng bằng phương pháp học máy - phương pháp được sử dụng rộng rãi trong đánh giá rủi ro tín dụng của người vay để giúp ngân hàng và người vay chống lại rủi ro tài chính một cách hiệu quả. Đặc biệt trong bối cảnh dịch bệnh hoặc những chủng bệnh mới có khả năng bùng phát bất cứ lúc nào ở toàn cầu hiện nay tương tự dịch COVID 19, suy thoái kinh tế toàn cầu là điều hiển nhiên, rủi ro tài chính đang gia tăng và thị trường tài chính bị tác động nghiêm trọng càng làm nổi bật tính hiệu quả và chính xác của các phương pháp học máy. Theo báo cáo năm 2019 của McKinsey, học máy có thể giảm 10% tổn thất tín dụng của ngân hàng và giảm thời gian ra quyết định tín dụng từ 25% đến 50%. Zestfinance - một trong những công ty đầu tiên áp dụng các phương pháp học máy để đánh giá rủi ro tín dụng, nhận thấy rằng hiệu suất của các mô hình học máy cao hơn 40% so với các mô hình đánh giá tín dụng truyền thống. Vì vậy, sâu xa hơn phương pháp nghiên cứu và kết luận của nhóm có thể áp dụng cho các lĩnh vực sau:

- Đầu tiên, các phương pháp và kết luận của nhóm có thể giúp các ngân hàng thiết lập mô hình chấm điểm tín dụng đối với doanh nghiệp vay vốn. Cách tiếp cận của nhóm hiện có thể giúp dự đoán điểm tín dụng và khả năng vỡ nợ của bên vay kết hợp với tình trạng tài sản của bên vay, hồ sơ tín dụng, khả năng sinh lời của hoạt động kinh doanh hiện tại, tốc độ tăng trưởng kinh doanh, vòng quay vốn lưu động, triển vọng phát triển ngành, tình hình hoạt động của các doanh nghiệp trực thuộc và dữ liệu kiểm toán công thương nghiệp và cục thuế. Trước tiên, chúng ta có thể sử dụng chương trình học máy để phân tích và tiền xử lý sơ bộ dữ liệu liên quan của doanh nghiệp đi vay, sau đó sử dụng phương pháp Random Forest hoặc thuật toán XGBoost để thực hiện lựa chọn tính năng, phân tách dữ liệu, xây dựng mô hình đào tạo học máy và sử dụng học máy mô hình để hoàn thành dự đoán dữ liệu. Ở giai đoạn cuối, để đo lường hiệu quả của mô hình, chúng ta có thể sử dụng các chỉ số đánh giá khác nhau, chẳng hạn như độ chính xác *(precision, recall, f1 score),* cohen-kappa, độ chính xác và khả năng thu hồi *(accuracy)*, để đánh giá hiệu quả dự đoán. Ngoài ra, chúng ta cũng có thể sử dụng các phương pháp học máy để đánh giá tình trạng tín dụng của những người vay cá nhân. Kết

hợp với dữ liệu về tài sản và hồ sơ tín dụng của từng người vay, phương tiện truyền thông xã hội như tham gia Twitter hoặc Facebook, thanh toán mạng, nền tảng cá nhân, văn hóa và tôn giáo, mạng xã hội và các đặc điểm nhân khẩu học như giới tính, tuổi tác, nghề nghiệp, giáo dục, có khả năng cung cấp dữ liệu sẽ được sử dụng để dự đoán điểm tín dụng và khả năng vỡ nợ của người đi vay.

- Thứ hai, phương pháp nghiên cứu và kết luận của nhóm cũng mang lại giá trị tham khảo tốt cho các cơ quan quản lý tài chính. Đầu tiên, các cơ quan quản lý có thể sử dụng dữ liệu lớn của doanh nghiệp để dự đoán rủi ro tài chính khu vực và sau đó thực hiện các biện pháp điều tiết, chẳng hạn như giảm hoặc tăng vốn lưu động, tăng hoặc giảm tỷ lệ dự trữ ngân hàng và tăng hoặc giảm lãi suất cho vay của ngân hàng, để đạt được mục đích quản lý rủi ro tài chính.
- Thứ ba, các cơ quan quản lý có thể sử dụng các phương pháp học máy để xác định gian lận tài chính và đưa ra dữ liệu giám sát rủi ro cho người cho vay thông qua các cảnh báo rủi ro. Cụ thể, các cơ quan quản lý có thể trích xuất dữ liệu giao dịch của thẻ tín dụng, sử dụng các phương pháp học máy để đào tạo và kiểm tra ngược, trích xuất các đặc điểm chính của gian lận tài chính và phân biệt các đặc điểm này với các giao dịch thông thường. Cơ chế nhận dạng này giúp cơ quan quản lý xác định hiệu quả hành vi gian lận tài chính, đặc biệt là những hành vi có điều kiện kinh doanh kém lợi dụng việc phòng ngừa và kiểm soát dịch bệnh *(nếu có)* để lừa đảo các khoản trợ cấp của chính phủ.

Ngoài ra, theo kết luận nghiên cứu đề tài, nhóm tin rằng chính phủ có thể quan tâm nhiều hơn đến khách hàng tín dụng cá nhân theo những cách sau để giảm thiểu thảm họa tài chính và phục hồi từ chúng:

- Thứ nhất, xác định nhu cầu vay và khả năng trả nợ của bên vay. Đối với người vay cá nhân có nhu cầu vay nhưng không vay được tiền kịp thời, cần đánh giá lại triển vọng phát triển nghề nghiệp và mạng lưới quan hệ xã hội của người vay bằng phương pháp học máy và quan tâm nhiều hơn đến tương lai hơn là hồ sơ tín dụng, tài sản, tài sản thế chấp và các thông tin quá khứ khác. Đối với những người vay có triển vọng phát triển tốt, chính phủ có thể ban hành các chính sách trợ cấp để giúp người vay vượt qua khó khăn.

- Thứ hai, các đề xuất đầu tư nên được đưa ra cho từng người vay theo mức độ chấp nhận rủi ro và khả năng trả nợ của những người vay khác nhau. Các kênh đầu tư như thị trường chứng khoán, thị trường trái phiếu, ngân hàng và tài chính internet có các dạng rủi ro, điều khoản, chi phí và lợi nhuận khác nhau.
- Thứ ba, các tổ chức tài chính được khuyến khích cấp các khoản vay cho những người vay cá nhân có xếp hạng tín dụng cao dưới hình thức trợ cấp và chính sách ưu đãi cho các tổ chức tài chính

Nhìn chung, các phương pháp nghiên cứu và kết luận của nhóm có thể được sử dụng để tham khảo bởi các ngân hàng, cơ quan quản lý tài chính, chính phủ, người đi vay doanh nghiệp và người vay cá nhân.

6.2. Hạn chế

Nghiên cứu được tiến hành còn những hạn chế ảnh hưởng đến kết quả nghiên cứu:

Tập dữ liệu bao gồm ba tệp Excel khác nhau bao gồm:

- 'loan-defaulter/application_data.csv'
- 'loan-defaulter/previous_application.csv'
- 'loan-defaulter/columns_description.csv'

Qua đó, trong 3 tập dữ liệu chỉ có tập dữ liệu ('loan-defaulter/application_data.csv') được sử dụng. Bộ dữ liệu 'loan-defaulter/previous_application.csv' là không liên quan. Bộ dữ liệu được sử dụng trong nghiên cứu có nhiều biến số quan trọng, được sử dụng trong các đơn xin vay vốn của các ngân hàng. Đổi lại, nhiều biến phải bị loại bỏ do mô tả không rõ ràng hoặc bị thiếu. Ngoài ra, theo như đề cập rằng tập dữ liệu là tập dữ liệu thực tế nhưng nguồn gốc hoặc quốc gia không được đề cập đến.

Thêm vào đó là tập dữ liệu không bao gồm tất cả các biến quan trọng. Một biến số rất thú vị là điểm tín dụng. Điểm tín dụng có thể là một số từ 0 đến 5, nghĩa là nếu người nộp đơn có điểm tín dụng bằng 0, thì người đó đã gặp vấn đề với các khoản vay hoặc thấu chi (Overdraft) tài khoản trước đó. Nếu điểm là 5, người nộp đơn đã thanh toán mọi khoản thanh toán đúng hạn và đã duy trì các tài khoản phù hợp. Một hạn chế khác là khả năng tính toán của máy tính mà nghiên cứu được thực hiện. Việc lấy mẫu quá

mức tập dữ liệu cho Random Forest khiến dữ liệu quá lớn và không thể thực thi được dẫn đến nhóm phải lấy mẫu đại diện và dùng xác suất để tính toán *(chia mẫu dữ liệu thành 5 phần nếu phần 1 là mẫu chuẩn dùng để tính toán thì phần 2,3,4,5 sẽ là phần train của model và xoay vòng tương tự)* để được kết quả chính xác nhất.

6.3. Hướng phát triển

Tiềm năng trong tương lai cho dự án này sẽ là sự phát triển hơn nữa của mô hình bằng cách phân tích sâu hơn về các biến được sử dụng trong các mô hình cũng như tạo mới các biến để đưa ra dự đoán tốt hơn. Đối với nghiên cứu trong tương lai, dự đoán vỡ nợ khoản vay nên được thực hiện cho một bộ dữ liệu tài sản thế chấp. Các quy tắc và hạn chế đối với các khoản thế chấp khác và nghiêm ngặt hơn so với các khoản vay quay vòng. Các khoản vay quay vòng không cần tài sản thế chấp trong xã hội, vì vậy tỷ lệ cho vay trên giá trị không áp dụng cho chúng. Tỷ lệ cho vay trên giá trị được tính dựa trên giá của căn nhà đã mua hoặc tài sản thế chấp khác. Một bộ dữ liệu thế chấp sẽ cho phép nghiên cứu thêm tầm quan trọng của hạn chế mới. Nghiên cứu trong tương lai nên giải quyết làm thế nào để duy trì những giá trị bị xóa để đơn giản hóa trong nghiên cứu nhưng phải lọc kỹ những biến số có khả năng ảnh hưởng lớn đến kết quả nghiên cứu, để mô hình được chuẩn xác nhất. Đồng thời, dựa vào tỷ lệ của nghiên cứu về khả năng vỡ nợ, các doanh nghiệp có thể đưa ra quyết định cho các nhân vay nợ hay không không từ ban đầu dựa trên các mô hình phân tích từ chỉ số trước đó *(ví dụ chỉ số nghiên cứu một biến số nào cao dẫn đến vỡ nợ thì từ khi người đi vay đưa hồ sơ vay vốn, doanh nghiệp có thể dựa vào chỉ số của họ và đưa ra quyết định kịp thời và chính xác)*, dẫn đến giảm rủi ro tài chính không mong muốn từ thời điểm ban đầu.

Tài liệu tham khảo

[1] Alomari, Z., & Fingerman, D. J. N. Z. J. o. C.-H. I. (2017). Loan Default Prediction and Identification of Interesting Relations between Attributes of Peer-to-Peer Loan Applications. *2*(2), 1-21.

[2] Anand, M., Velu, A., Whig, P. J. J. o. C. S., & Engineering. (2022). Prediction of loan behaviour with machine learning models for secure banking. *3*(1), 1-13.

[3] Aslam, U., Tariq Aziz, H. I., Sohail, A., Batcha, N. K. J. J. o. C., & Nanoscience, T. (2019). An empirical study on loan default prediction models. *16*(8), 3483-3488.

[4] Boyapati, M., & Aygun, R. (2023). *Default Prediction on Commercial Credit Big Data Using Graph-based Variable Clustering.* Paper presented at the 2023 IEEE 17th International Conference on Semantic Computing (ICSC).

[5] Coşer, A., Maer-matei, M. M., Albu, C. J. E. C., Studies, E. C., & Research. (2019). PREDICTIVE MODELS FOR LOAN DEFAULT RISK ASSESSMENT. *53*(2).

[6] Chen, Y.-Q., Zhang, J., & Ng, W. W. (2018). *Loan default prediction using diversified sensitivity undersampling.* Paper presented at the 2018 International Conference on Machine Learning and Cybernetics (ICMLC).

[7] Datkhile, A., Chandak, K., Bhandari, S., Gajare, H., & Karyakarte, M. J. I. (2020). Statistical Modelling on Loan Default Prediction Using Different Models. *3*(3), 3-5.

[8] Eweoya, I., Adebiyi, A., Azeta, A., & Azeta, A. E. (2019). *Fraud prediction in bank loan administration using decision tree.* Paper presented at the Journal of Physics: Conference Series.

[9] Guo, W., & Zhou, Z. Z. J. J. o. F. (2022). A comparative study of combining tree-based feature selection methods and classifiers in personal loan default prediction. *41*(6), 1248-1313.

[10] Jiang, W. J. I. T. o. I., & Systems. (2022). Loan Default Prediction with Deep Learning and Muddling Label Regularization. *105*(7), 1340-1342.

[11] Li, W., Ding, S., Chen, Y., & Yang, S. J. I. A. (2018). Heterogeneous ensemble for default prediction of peer-to-peer lending in China. *6*, 54396-54406.

[12] Li, X., Ergu, D., Zhang, D., Qiu, D., Cai, Y., & Ma, B. J. P. C. S. (2022). Prediction of loan default based on multi-model fusion. *199*, 757-764.

[13] Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). *Loan default prediction using decision trees and random forest: A comparative study.* Paper presented at the IOP Conference Series: Materials Science and Engineering.

[14] Netzer, O., Lemaire, A., & Herzenstein, M. J. J. o. M. R. (2019). When words sweat: Identifying signals for loan default in the text of loan applications. *56*(6), 960-980.

[15] Padimi, V., Venkata, S., Devarani, D. J. J. o. N., & Systems, F. (2022). Applying Machine Learning Techniques To Maximize The Performance of Loan Default Prediction. *2*(2), 44-56.

[16] Song, Y., Wang, Y., Ye, X., Zaretzki, R., & Liu, C. J. I. S. (2023). Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme. *629*, 599-617.

[17] Stevenson, M., Mues, C., & Bravo, C. J. E. J. o. O. R. (2021). The value of text for small business default prediction: A deep learning approach. *295*(2), 758-771.

[18] Wang, H., Chen, W., & Da, F. J. P. C. S. (2022). Zhima Credit Score in Default Prediction for Personal Loans. *199*, 1478-1482.

[19] Zhang, L., Wang, J., & Liu, Z. J. E. S. w. A. (2023). What should lenders be more concerned about? Developing a profit-driven loan default prediction model. *213*, 118938.

[20] Zhu, Q., Ding, W., Xiang, M., Hu, M., Zhang, N. J. I. J. o. D. W., & Mining. (2023). Loan Default Prediction Based on Convolutional Neural Network and LightGBM. *19*(1), 1-16.

[21] Ahmed, M. I. and P. R. Rajaleximi (2019). "An empirical study on credit scoring and credit scorecard for financial institutions." International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 8(7): 2278-1323.

[22] Aslam, U., et al. (2019). "An empirical study on loan default prediction models." Journal of Computational and Theoretical Nanoscience 16(8): 3483-3488.

[23] Russell, S. and P. Norvig (1995). "Artificial intelligence: A modern approach prentice-hall." Englewood cliffs.

[24] Russell, S. and P. Norvig (1995). "A modern, agent-oriented approach to introductory artificial intelligence." Acm Sigart Bulletin 6(2): 24-26.

[25] Shoumo, S. Z. H., et al. (2019). Application of machine learning in credit risk assessment: a prelude to smart banking. TENCON 2019-2019 IEEE Region 10 Conference (TENCON), IEEE.

[26] Zhu, L., et al. (2019). "A study on predicting loan default based on the random forest algorithm." Procedia Computer Science 162: 503-513.

[27] Odegua, R. (2020). "Predicting bank loan default with extreme gradient boosting." arXiv preprint arXiv:2002.02011.

[28] Gautam, K., et al. (2020). "Loan Prediction using Decision Tree and Random Forest." International Research Journal of Engineering and Technology (IRJET) 7(08): 853-856.

[29] Natekin, A. and A. Knoll (2013). "Gradient boosting machines, a tutorial." Frontiers in neurorobotics 7: 21.

[30] Manglani, R. and A. Bokhare (2021). Logistic Regression Model for Loan Prediction: A Machine Learning Approach. 2021 Emerging Trends in Industry 4.0 (ETI 4.0), IEEE.

[31] Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. Statistics and Computing, 21, 137-146.

[32] Rau, G., & Shih, Y.-S. (2021). Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. Journal of English for academic purposes, 53, 101026.