BỘ NÔNG NGHIỆP VÀ MÔI TRƯỜNG

PHÂN HIỆU TRƯỜNG ĐẠI HỌC THỦY LỢI

BỘ MÔN CÔNG NGHỆ THÔNG TIN

BÁO CÁO MÔN HỌC

BỘ MÔN: DATA MINING

TÊN ĐỀ TÀI:

PHÂN LOẠI HỒ SƠ GIAN LẬN

Giảng viên	**Cô Vũ Thị Hạnh**
Sinh viên thực hiện	**Nguyễn Phước Toàn** **Lê Văn Quang** **Nguyễn Hoàng Tuấn Anh**
Lớp	**S26-65TTNT**

TP. Hồ Chí Minh, ngày … tháng … năm 2026

LỜI CẢM ƠN

Lời đầu tiên, nhóm em xin gửi lời cảm ơn chân thành đến Khoa Công nghệ Thông tin đã đưa môn học Data Mining vào chương trình giảng dạy. Đây là một môn học có ý nghĩa thiết thực, giúp nhóm em tiếp cận và vận dụng các kiến thức về xử lý dữ liệu, phân tích dữ liệu và ứng dụng trí tuệ nhân tạo vào các bài toán thực tế.

Trong quá trình học tập và thực hiện đề tài "Phân loại hồ sơ gian lận", nhóm em đã nhận được sự quan tâm, hướng dẫn tận tình của các thầy, cô trong Bộ môn Công nghệ Thông tin. Đặc biệt, nhóm em xin bày tỏ lòng biết ơn sâu sắc đến Cô Vũ Thị Hạnh, người đã trực tiếp hướng dẫn và định hướng cho nhóm em trong suốt quá trình thực hiện bài. Những góp ý và chỉ dẫn của Cô đã giúp nhóm em hiểu rõ hơn về bài toán phân loại ảnh, quy trình xử lý dữ liệu lớn cũng như cách xây dựng và đánh giá các mô hình học sâu.

Trong quá trình thực hiện đề tài, do kiến thức và kinh nghiệm thực tế của nhóm em còn hạn chế, bài làm khó tránh khỏi những thiếu sót. Tuy nhiên, nhóm em đã cố gắng vận dụng tối đa những kiến thức đã học để xây dựng mô hình, thực nghiệm và phân tích kết quả một cách nghiêm túc. Nhóm em rất mong nhận được những nhận xét và góp ý từ Cô để bài báo cáo được hoàn thiện hơn và trở thành nền tảng kiến thức hữu ích cho quá trình học tập và công việc sau này.

Nhóm em xin chân thành cảm ơn

MỤC LỤC

MỞ ĐẦU

Trong những năm gần đây, sự phát triển mạnh mẽ của các dịch vụ tài chính số và ngân hàng trực tuyến đã mang lại nhiều tiện ích cho người dùng. Tuy nhiên, song song với đó là sự gia tăng nhanh chóng của các hành vi gian lận, đặc biệt là gian lận trong quá trình mở tài khoản ngân hàng. Các đối tượng gian lận có thể lợi dụng thông tin giả mạo, thiết bị không đáng tin cậy hoặc hành vi bất thường để tạo ra các tài khoản phục vụ cho mục đích rửa tiền, lừa đảo hoặc các hoạt động phi pháp khác.

Việc phát hiện sớm các tài khoản gian lận đóng vai trò vô cùng quan trọng đối với các tổ chức tài chính, giúp giảm thiểu rủi ro, tổn thất kinh tế và bảo vệ uy tín của hệ thống ngân hàng. Trong bối cảnh đó, các phương pháp khai phá dữ liệu (Data Mining) và học máy (Machine Learning) ngày càng được ứng dụng rộng rãi nhằm tự động hóa quá trình phát hiện gian lận với độ chính xác cao.

Đề tài này tập trung nghiên cứu và xây dựng mô hình học máy để phát hiện gian lận khi mở tài khoản ngân hàng dựa trên dữ liệu thực tế, qua đó đánh giá hiệu quả của các phương pháp tiền xử lý, cân bằng dữ liệu và mô hình phân loại hiện đại.

CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

1.1 Lí do chọn đề tài

Trong bối cảnh chuyển đổi số mạnh mẽ của lĩnh vực tài chính – ngân hàng, các hình thức gian lận ngày càng trở nên tinh vi và khó phát hiện. Đặc biệt, gian lận trong quá trình **mở tài khoản ngân hàng trực tuyến** gây ra nhiều rủi ro nghiêm trọng, không chỉ làm tổn thất tài chính mà còn ảnh hưởng đến uy tín và độ tin cậy của các tổ chức tài chính.

Các hệ thống phát hiện gian lận truyền thống dựa trên luật (rule-based) thường không còn đáp ứng tốt do thiếu khả năng thích nghi với những hành vi gian lận mới. Trong khi đó, **khai phá dữ liệu (Data Mining)** và **học máy (Machine Learning)** cho phép phân tích dữ liệu lớn, phát hiện các mẫu tiềm ẩn và đưa ra quyết định chính xác hơn trong môi trường dữ liệu phức tạp và mất cân bằng.

Xuất phát từ thực tiễn đó, nhóm lựa chọn đề tài **"Phát hiện gian lận trong mở tài khoản ngân hàng bằng kỹ thuật khai phá dữ liệu"** nhằm áp dụng các thuật toán học máy hiện đại trên bộ dữ liệu thực tế, qua đó đánh giá khả năng phát hiện gian lận và đề xuất giải pháp hỗ trợ ra quyết định cho hệ thống ngân hàng.

1.2 Mục tiêu nghiên cứu

Mục tiêu của nghiên cứu này là xây dựng và đánh giá một mô hình học máy có khả năng phát hiện các hồ sơ mở tài khoản ngân hàng gian lận dựa trên dữ liệu hành vi và thông tin định danh của khách hàng. Bài toán được mô hình hóa dưới dạng bài toán phân loại nhị phân, trong đó mỗi hồ sơ được gán nhãn *gian lận* hoặc *không gian lận*.

Cụ thể, nghiên cứu hướng tới các mục tiêu sau:

- Phân tích và hiểu rõ bộ dữ liệu gian lận ngân hàng, bao gồm đặc điểm phân phối dữ liệu, mức độ mất cân bằng giữa các lớp và sự biến động của tỷ lệ gian lận theo thời gian.
- Xây dựng quy trình tiền xử lý dữ liệu phù hợp, bao gồm xử lý giá trị thiếu, mã hóa biến phân loại, chuẩn hóa dữ liệu và lựa chọn đặc trưng, nhằm đảm bảo dữ liệu đầu vào phù hợp với các thuật toán học máy.
- Ứng dụng các mô hình học máy dựa trên cây quyết định, đặc biệt là LightGBM, để khai thác các mối quan hệ phi tuyến và tương tác phức tạp giữa các đặc trưng trong bài toán phát hiện gian lận.
- Giải quyết vấn đề mất cân bằng dữ liệu, thông qua việc điều chỉnh trọng số lớp hoặc ngưỡng phân loại, nhằm nâng cao khả năng phát hiện các trường hợp gian lận hiếm gặp.
- Đánh giá hiệu quả mô hình bằng các chỉ số phù hợp với bài toán thực tế, bao gồm Recall, Precision, F1-score và AUC-ROC, thay vì chỉ dựa trên độ chính xác (Accuracy).
- Phân tích kết quả và khả năng ứng dụng thực tiễn của mô hình, từ đó đề xuất hướng cải thiện và mở rộng trong các nghiên cứu tiếp theo.

1.3 Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài bao gồm:

- Dữ liệu nghiên cứu:
  Bộ dữ liệu Bank Account Fraud Dataset (NeurIPS 2022), bao gồm các thông tin liên quan đến nhân khẩu học, hành vi giao dịch, thiết bị và danh tính của người dùng trong quá trình mở tài khoản ngân hàng. Biến mục tiêu là fraud_bool, biểu thị trạng thái gian lận hoặc không gian lận.
- Đối tượng phân tích:
  Các hồ sơ đăng ký mở tài khoản ngân hàng, được phân loại thành hai nhóm:

    - Hồ sơ hợp lệ (không gian lận)
    - Hồ sơ gian lận
- Phương pháp nghiên cứu:

  Áp dụng các kỹ thuật khai phá dữ liệu và học máy có giám sát, đặc biệt là các mô hình dựa trên cây tăng cường (Gradient Boosting) như LightGBM, để xây dựng hệ thống phát hiện gian lận hiệu quả.

CHƯƠNG 2 : CƠ SỞ LÝ THUYẾT LIÊN QUAN

2.1 Khai phá dữ liệu

Khai phá dữ liệu (Data Mining) là quá trình khám phá các mẫu, tri thức tiềm ẩn và các mối quan hệ có ý nghĩa từ những tập dữ liệu lớn. Quá trình này kết hợp nhiều lĩnh vực như thống kê, học máy, cơ sở dữ liệu và trí tuệ nhân tạo nhằm hỗ trợ ra quyết định.

Một quy trình khai phá dữ liệu điển hình bao gồm các bước:

1. Thu thập và hiểu dữ liệu
2. Tiền xử lý dữ liệu
3. Xây dựng mô hình
4. Đánh giá mô hình
5. Triển khai và ứng dụng

Trong đề tài này, khai phá dữ liệu được áp dụng để phân tích các hồ sơ mở tài khoản ngân hàng, từ đó phát hiện các trường hợp có dấu hiệu gian lận.

2.2. Tiền xử lý dữ liệu (Data Preprocessing)

Tiền xử lý dữ liệu là bước quan trọng nhằm đảm bảo chất lượng dữ liệu đầu vào cho mô hình học máy. Các kỹ thuật tiền xử lý chính được sử dụng trong đề tài bao gồm:

2.2.1. Mã hóa biến phân loại (One-Hot Encoding)

Các biến phân loại (categorical features) không thể được sử dụng trực tiếp trong các mô hình học máy. Do đó, kỹ thuật One-Hot Encoding được sử dụng để chuyển đổi mỗi giá trị phân loại thành một vector nhị phân, giúp mô hình học được thông tin từ các biến này.

2.2.2. Chuẩn hóa dữ liệu (Min-Max Scaling)

Chuẩn hóa dữ liệu giúp đưa các biến số về cùng một khoảng giá trị, thường là [0,1][0,1][0,1]. Điều này giúp quá trình huấn luyện mô hình ổn định hơn và tránh việc các biến có giá trị lớn chi phối mô hình.

2.2.3. Xử lý dữ liệu mất cân bằng

Do tỷ lệ gian lận rất thấp, dữ liệu có xu hướng mất cân bằng nghiêm trọng. Để khắc phục vấn đề này, các kỹ thuật cân bằng dữ liệu như:

- Undersampling (NearMiss)
- Oversampling
- Điều chỉnh trọng số lớp (class_weight)

được áp dụng nhằm giúp mô hình học tốt hơn đặc trưng của lớp thiểu số (gian lận).

2.4. Học máy có giám sát (Supervised Learning)

Học máy có giám sát là phương pháp học trong đó mô hình được huấn luyện trên dữ liệu đã có nhãn. Mục tiêu là học được mối quan hệ giữa dữ liệu đầu vào và nhãn đầu ra để dự đoán cho các dữ liệu mới.

Trong đề tài này, học máy có giám sát được sử dụng để xây dựng mô hình phân loại gian lận dựa trên các đặc trưng của hồ sơ mở tài khoản.

2.5. Thuật toán LightGBM

LightGBM (Light Gradient Boosting Machine) là một thuật toán học máy dựa trên kỹ thuật **Gradient Boosting Decision Tree (GBDT)**, được phát triển bởi Microsoft. LightGBM được thiết kế để xử lý dữ liệu lớn với hiệu suất cao và độ chính xác tốt.

2.5.1. Nguyên lý hoạt động

LightGBM xây dựng mô hình bằng cách:

- Huấn luyện nhiều cây quyết định theo từng vòng lặp
- Mỗi cây mới được xây dựng để khắc phục sai số của các cây trước đó
- Các cây được kết hợp để tạo ra mô hình cuối cùng

Khác với các thuật toán GBDT truyền thống, LightGBM sử dụng chiến lược **leaf-wise growth** thay vì level-wise, giúp giảm sai số nhanh hơn.

2.5.2. Ưu điểm của LightGBM

- Huấn luyện nhanh và tiết kiệm bộ nhớ
- Xử lý tốt dữ liệu có kích thước lớn
- Hiệu quả cao đối với dữ liệu mất cân bằng
- Hỗ trợ điều chỉnh ngưỡng dự đoán (threshold) linh hoạt

2.6. Thuật toán XGBoost

XGBoost (Extreme Gradient Boosting) là một thuật toán học máy mạnh mẽ dựa trên kỹ thuật **Gradient Boosting**, được thiết kế với mục tiêu tối ưu hóa hiệu năng và khả năng mở rộng. XGBoost đã được sử dụng rộng rãi trong nhiều cuộc thi khoa học dữ liệu và các hệ thống thực tế nhờ khả năng đạt độ chính xác cao.

2.6.1. Nguyên lý hoạt động

XGBoost xây dựng mô hình bằng cách:

- Huấn luyện tuần tự các cây quyết định
- Mỗi cây mới tập trung vào việc giảm hàm mất mát (loss function) của toàn bộ mô hình

- Sử dụng đạo hàm bậc nhất và bậc hai của hàm mất mát để tối ưu quá trình huấn luyện

Thuật toán bổ sung các cơ chế regularization nhằm kiểm soát độ phức tạp của mô hình, giúp hạn chế hiện tượng quá khớp (overfitting).

2.6.2. Ưu điểm của XGBoost

- Hiệu quả cao và độ chính xác tốt
- Hỗ trợ regularization mạnh mẽ (L1, L2)
- Xử lý tốt dữ liệu phi tuyến
- Có khả năng mở rộng cho tập dữ liệu lớn
- Phù hợp với các bài toán phân loại nhị phân như phát hiện gian lận

2.7. Thuật toán CatBoost

CatBoost (Categorical Boosting) là một thuật toán Gradient Boosting được phát triển bởi Yandex, đặc biệt tối ưu cho dữ liệu chứa nhiều biến phân loại.

2.7.1. Nguyên lý hoạt động

CatBoost sử dụng kỹ thuật **Ordered Boosting** để:

- Giảm hiện tượng rò rỉ dữ liệu (target leakage)
- Xử lý trực tiếp các biến phân loại mà không cần One-Hot Encoding thủ công

Thuật toán mã hóa biến phân loại dựa trên thống kê có điều kiện, giúp mô hình học được mối quan hệ phức tạp giữa các biến và nhãn mục tiêu.

2.7.2. Ưu điểm của CatBoost

- Xử lý biến phân loại hiệu quả
- Ít cần tiền xử lý dữ liệu

- Giảm nguy cơ overfitting
- Hoạt động ổn định trên dữ liệu mất cân bằng
- Phù hợp với các bài toán gian lận tài chính

2.6. Đánh giá mô hình phân loại

Việc đánh giá mô hình trong bài toán gian lận không chỉ dựa vào độ chính xác (Accuracy) mà còn cần xem xét các chỉ số khác:

- Precision: Tỷ lệ dự đoán gian lận đúng trên tổng số dự đoán gian lận
- Recall: Tỷ lệ phát hiện đúng gian lận trên tổng số gian lận thực tế
- F1-score: Trung bình điều hòa giữa Precision và Recall
- ROC Curve và AUC: Đánh giá khả năng phân biệt giữa hai lớp
- Confusion Matrix: Phân tích chi tiết các trường hợp dự đoán đúng và sai

Đặc biệt, trong đề tài này, mô hình được đánh giá tại ngưỡng quyết định tối ưu dựa trên False Positive Rate (FPR) nhằm đảm bảo hiệu quả phát hiện gian lận trong thực tế.

2.7. Điều chỉnh ngưỡng quyết định (Threshold Tuning)

Thông thường, các mô hình phân loại sử dụng ngưỡng mặc định là 0.5 để đưa ra dự đoán. Tuy nhiên, trong bài toán phát hiện gian lận, việc điều chỉnh ngưỡng quyết định là cần thiết để cân bằng giữa Precision và Recall.

Bằng cách lựa chọn ngưỡng dựa trên ROC Curve hoặc yêu cầu cụ thể về FPR, mô hình có thể đạt hiệu quả tốt hơn trong việc phát hiện gian lận, phù hợp với yêu cầu thực tế của hệ thống ngân hàng.

CHƯƠNG 3: MÔ TẢ DỮ LIỆU VÀ CÁC BƯỚC TIỀN XỬ LÝ

3.1 Mô tả dữ liệu

Trong nghiên cứu này, nhóm sử dụng tập dữ liệu Bank Account Fraud Dataset, bao gồm các thông tin liên quan đến hồ sơ khách hàng, hành vi giao dịch và thiết bị truy cập. Mỗi bản ghi đại diện cho một yêu cầu mở tài khoản hoặc giao dịch, được gán nhãn nhị phân nhằm xác định hành vi gian lận (Fraud) hoặc bình thường (Non-Fraud).

Biến mục tiêu trong bài toán là:

- fraud_bool:
    - Giá trị 1: giao dịch gian lận
    - Giá trị 0: giao dịch hợp lệ

Dữ liệu bao gồm các nhóm đặc trưng chính:

- Đặc trưng nhân khẩu học: tuổi khách hàng, thời gian cư trú, lịch sử ngân hàng
- Đặc trưng hành vi giao dịch: số tiền dự kiến, thời lượng phiên truy cập
- Đặc trưng thiết bị và phiên: số lượng email liên kết với thiết bị, thời gian phiên
- Đặc trưng phân loại: quốc gia, loại thiết bị, phương thức truy cập

Tập dữ liệu có đặc điểm mất cân bằng nghiêm trọng, trong đó tỷ lệ gian lận chiếm một phần rất nhỏ so với các giao dịch bình thường, gây khó khăn cho việc huấn luyện mô hình học máy.

3.2 Các bước tiền xử lý

3.2.1. Model LightGBM

- **Chia tập dữ liệu theo thời gian (Temporal Split):** Do đặc thù của dữ liệu gian lận tài chính thường thay đổi hành vi theo thời gian (concept drift), việc chia dữ

liệu ngẫu nhiên có thể gây ra hiện tượng rò rỉ dữ liệu (data leakage). Vì vậy, nhóm áp dụng phương pháp chia tách dựa trên tháng giao dịch:

  - Tập Huấn luyện (Training Set): Bao gồm các giao dịch từ tháng 0 đến tháng 5, dùng để mô hình học các quy luật.
  - Tập Kiểm thử (Test Set): Bao gồm các giao dịch từ tháng 6 trở đi, dùng để đánh giá khả năng dự báo trên dữ liệu tương lai.

- **Xử lý giá trị khuyết thiếu (Handling Missing Values):** Trong bộ dữ liệu, các giá trị thiếu được mã hóa dưới dạng số -1. Nhóm xác định các cột chứa giá trị này gồm: prev_address_months_count, current_address_months_count, bank_months_count, session_length_in_minutes, và device_distinct_emails_8w.
  - Phương pháp: Thay thế các giá trị -1 bằng giá trị trung vị (median) của chính cột đó.
  - Mục đích: Việc dùng trung vị giúp dữ liệu ít bị ảnh hưởng bởi các giá trị ngoại lai (outliers) hơn so với giá trị trung bình.
- **Kỹ thuật tạo đặc trưng mới (Feature Engineering):** Để tăng cường khả năng phát hiện gian lận, nhóm đã xây dựng thêm các đặc trưng phái sinh dựa trên dữ liệu gốc:
  - Velocity (Tốc độ giao dịch): Được tính bằng số tiền giao dịch chia cho thời gian phiên đăng nhập. Đặc trưng này giúp phát hiện hành vi tẩu tán tiền nhanh của kẻ gian.
  - Log Amount (Biến đổi Log): Áp dụng hàm logarit cho số tiền giao dịch để giảm độ lệch của phân phối dữ liệu, giúp mô hình hội tụ tốt hơn.
  - Age-Bank Interaction (Tương tác Tuổi - Thâm niên): Kết hợp giữa tuổi khách hàng và thời gian mở tài khoản để đánh giá mức độ tin cậy của hồ sơ.
- **Mã hóa và Chuẩn hóa dữ liệu (Encoding & Scaling):** Dữ liệu được chuyển đổi về dạng số học đồng nhất để phục vụ tính toán:
  - Mã hóa biến phân loại: Sử dụng One-Hot Encoding để chuyển đổi các biến định danh (như loại thanh toán, tình trạng nghề nghiệp) thành các vector nhị phân.

    - Chuẩn hóa biến số: Sử dụng RobustScaler thay vì StandardScaler. Phương pháp này sử dụng khoảng tứ phân vị (IQR) để chuẩn hóa, giúp giảm thiểu tác động tiêu cực của các giao dịch có giá trị đột biến (ngoại lai).
- **Lựa chọn đặc trưng (Feature Selection):** Nhóm áp dụng kỹ thuật loại bỏ đặc trưng dựa trên phương sai (Variance Threshold). Các cột dữ liệu có phương sai bằng 0 (chứa cùng một giá trị cho tất cả các mẫu) sẽ bị loại bỏ vì chúng không mang lại thông tin phân loại hữu ích cho mô hình.
- **Cân bằng dữ liệu (Data Balancing):** Dữ liệu gian lận có tính chất mất cân bằng nghiêm trọng (imbalanced), khiến mô hình dễ thiên vị lớp đa số.
    - Phương pháp: Sử dụng Random Under-sampling trên tập huấn luyện.
    - Tỷ lệ áp dụng: Giữ nguyên toàn bộ mẫu gian lận (Fraud) và giảm bớt mẫu bình thường (Non-Fraud) để đạt tỷ lệ tối ưu 1:2 (0.5).
    - Mục đích: Giúp mô hình tập trung học các đặc điểm của hành vi gian lận hiệu quả hơn mà không bị nhiễu bởi quá nhiều dữ liệu bình thường, đồng thời giảm thời gian huấn luyện.

3.2.2 Model CatBoost

Quy trình tiền xử lý cho mô hình này được thiết kế tinh gọn, tập trung vào việc kiến tạo đặc trưng và tối ưu hóa tham số nội bộ hơn là biến đổi dữ liệu đầu vào. Các bước cụ thể như sau:

- **Kiến tạo đặc trưng hành vi (Behavioral Feature Engineering):** Trước khi đưa vào mô hình, dữ liệu gốc được làm giàu thông qua bước apply_features. Nhóm nghiên cứu tập trung xây dựng các biến phái sinh nhằm làm nổi bật hành vi gian lận, bao gồm: tính toán tốc độ giao dịch (velocity), khoảng thời gian giữa các lần đăng nhập, và các cờ báo hiệu rủi ro (risk flags) dựa trên thiết bị hoặc địa chỉ IP. Bước này giúp chuyển đổi dữ liệu từ dạng thông tin tĩnh sang dạng thông tin hành vi động.

- **Chia tập dữ liệu phân tầng (Stratified Splitting):** Để đảm bảo tính đại diện của dữ liệu, nhóm sử dụng kỹ thuật lấy mẫu phân tầng (Stratified Sampling) khi chia tập Huấn luyện (Train) và Kiểm thử (Test) với tỷ lệ 80/20.
    - Mục đích: Đảm bảo tỷ lệ các ca gian lận (lớp thiểu số) được giữ nguyên trong cả hai tập dữ liệu, ngăn chặn hiện tượng tập Test quá ít mẫu gian lận dẫn đến đánh giá sai lệch.
- **Xử lý biến phân loại tự động (Native Categorical Handling):** Thay vì sử dụng các kỹ thuật mã hóa thủ công như One-Hot Encoding (gây bùng nổ số chiều dữ liệu) hay Label Encoding (gây nhiễu thứ tự), nhóm tận dụng cơ chế xử lý biến phân loại nội tại của CatBoost.
    - Thực hiện: Nhóm xác định danh sách các cột dữ liệu dạng chuỗi (Object types) và truyền trực tiếp vào tham số cat_features của mô hình.
    - Ưu điểm: CatBoost sử dụng kỹ thuật "Ordered Target Statistics" để mã hóa, giúp giữ lại tối đa thông tin của các biến phân loại mà không làm tăng kích thước bộ nhớ.
- **Xử lý mất cân bằng bằng trọng số (Cost-Sensitive Learning):** Thay vì can thiệp vào dữ liệu vật lý (như xóa bớt dữ liệu sạch bằng Undersampling hay sinh thêm dữ liệu giả bằng SMOTE), quy trình này sử dụng phương pháp "Phạt trọng số" (Class Weighting).
    - Cách tính: Hệ số phạt scale_pos_weight được tính toán động dựa trên tỷ lệ: Tổng số mẫu bình thường / Tổng số mẫu gian lận.
    - Cơ chế: Khi huấn luyện, mô hình sẽ bị phạt nặng hơn rất nhiều lần nếu dự báo sai một ca gian lận so với một ca bình thường. Điều này ép buộc mô hình phải "chú ý" đặc biệt đến các giao dịch gian lận mà không cần làm biến dạng phân phối dữ liệu gốc.
- **Chuẩn hóa dữ liệu (Scaling):** Đối với CatBoost (một thuật toán dựa trên cây quyết định), việc chuẩn hóa dữ liệu số (như Min-Max Scaling hay Standard Scaling) là không bắt buộc và không ảnh hưởng đáng kể đến độ chính xác. Do đó,

nhóm quyết định giữ nguyên giá trị gốc của các biến số để bảo toàn ý nghĩa thực tế của dữ liệu (ví dụ: giữ nguyên số tiền là USD thay vì chuyển về khoảng 0-1).

3.2.3 Model XGBoost

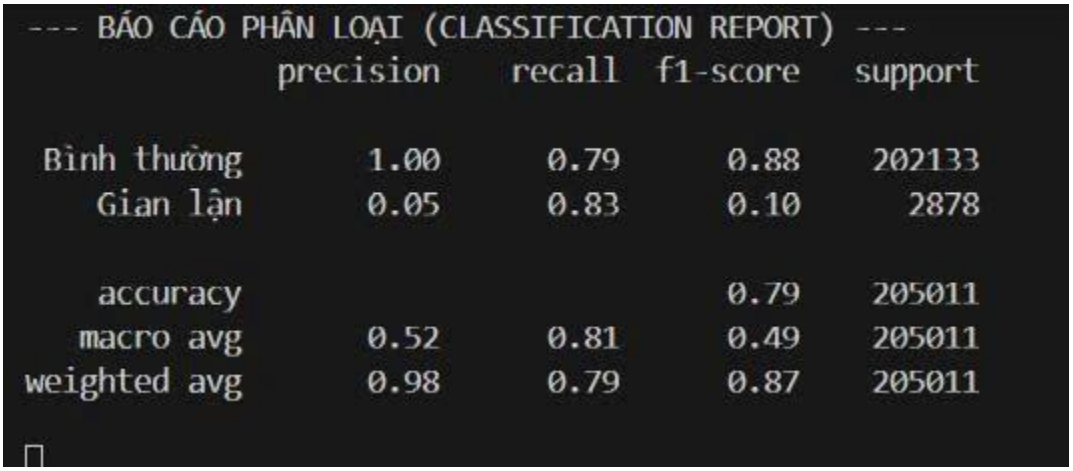
Để tối ưu hóa hiệu suất cho mô hình XGBoost và giải quyết vấn đề mất cân bằng dữ liệu nghiêm trọng, nhóm nghiên cứu đã áp dụng một quy trình tiền xử lý đặc thù, tập trung vào việc làm giàu dữ liệu (Feature Enrichment) và lọc nhiễu ngay từ đầu vào. Các bước thực hiện cụ thể như sau:

- **Cân bằng dữ liệu sơ cấp (Majority Downsampling):** Trước khi đưa vào huấn luyện, nhóm thực hiện giảm mẫu (Downsampling) đối với lớp đa số (giao dịch bình thường - Non-Fraud).
    - Tỷ lệ áp dụng: Thay vì giữ nguyên tỷ lệ gốc (1:100), nhóm đưa dữ liệu về tỷ lệ **1:20** (1 giao dịch gian lận tương ứng với 20 giao dịch bình thường).
    - Mục đích: Loại bỏ lượng lớn dữ liệu nhiễu, giúp mô hình tập trung học các mẫu khó và giảm tải tài nguyên tính toán, trong khi vẫn giữ lại đủ độ đa dạng của các giao dịch bình thường để tránh báo động giả (False Positive).
- **Chia tập dữ liệu (Stratified Splitting):** Sau khi cân bằng sơ bộ, dữ liệu được chia thành tập Huấn luyện (Train) và tập Kiểm thử (Test) theo tỷ lệ 80/20.
    - Phương pháp: Sử dụng kỹ thuật lấy mẫu phân tầng (Stratified Sampling) để đảm bảo tỷ lệ gian lận trong cả tập Train và tập Test là tương đồng nhau, giữ nguyên tính chất phân phối của dữ liệu sau khi cân bằng.
- **Làm sạch dữ liệu số (Numeric Cleaning):** Các cột dữ liệu số quan trọng (như velocity, days_since_request, income...) được ép kiểu dữ liệu (forcing numeric types) để xử lý các lỗi định dạng tiềm ẩn, đảm bảo tính toán thống kê chính xác ở các bước sau.

- **Kỹ thuật đặc trưng nâng cao (Risk Score Engineering):** Đây là bước quan trọng nhất trong quy trình. Thay vì để mô hình tự học từ dữ liệu thô, nhóm đã xây dựng hệ thống "Điểm rủi ro" (Risk Score) dựa trên các ngưỡng thống kê:
    - Tính toán ngưỡng (Thresholding): Tính toán các giá trị phân vị (Quantile) 5% và 95% trên tập Huấn luyện để xác định các hành vi bất thường (ví dụ: thời gian phiên quá ngắn, tốc độ rút tiền quá nhanh).
    - Tạo cờ báo động (Risk Flags): Tạo ra các biến nhị phân (Binary Flags) đánh dấu các hành vi khả nghi như: very_short_session (phiên đăng nhập cực ngắn), high_velocity_6h (tốc độ giao dịch cao trong 6h), many_emails_same_device (nhiều email trên một thiết bị).
    - Tổng hợp Risk Score: Cộng tổng các cờ báo động để tạo thành một đặc trưng mới là risk_score. Đặc trưng này đóng vai trò như một chỉ báo mạnh mẽ giúp mô hình phân tách rõ ràng giữa gian lận và bình thường.
- **Mã hóa và Chuẩn hóa (Encoding & Scaling):** Để đưa vào mô hình XGBoost, dữ liệu được chuyển đổi qua Pipeline xử lý tự động:
    - Chuẩn hóa biến số: Sử dụng **StandardScaler** để đưa các biến số về phân phối chuẩn (mean=0, variance=1), giúp thuật toán Gradient Descent hội tụ nhanh hơn.
    - Mã hóa biến phân loại: Sử dụng **One-Hot Encoding** với cơ chế xử lý biến lạ (handle_unknown='ignore') để chuyển đổi các biến định danh, đảm bảo mô hình hoạt động ổn định ngay cả khi gặp các danh mục mới trong tương lai.
- **Xử lý mất cân bằng trọng số (Class Weighting):** Bên cạnh việc Downsampling, nhóm kết hợp thêm tham số phạt scale_pos_weight trong cấu hình mô hình. Trọng số này được thiết lập bằng 20 (tương ứng với tỷ lệ mẫu 1:20), ép buộc mô hình phải "coi trọng" việc phân loại đúng một giao dịch gian lận gấp 20 lần so với một giao dịch bình thường.

CHƯƠNG 4: KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH

4.1 LightGBM



Accuracy = 0,79 (79%)

Mặc dù độ chính xác tổng thể tương đối cao, nhưng do dữ liệu mất cân bằng nghiêm trọng nên accuracy không phản ánh đầy đủ chất lượng mô hình trong việc phát hiện gian lận.

Đối với lớp Gian lận (1):

- Precision = 0,05 (5%)
- Recall = 0,83 (83%)
- F1-score = 0,10

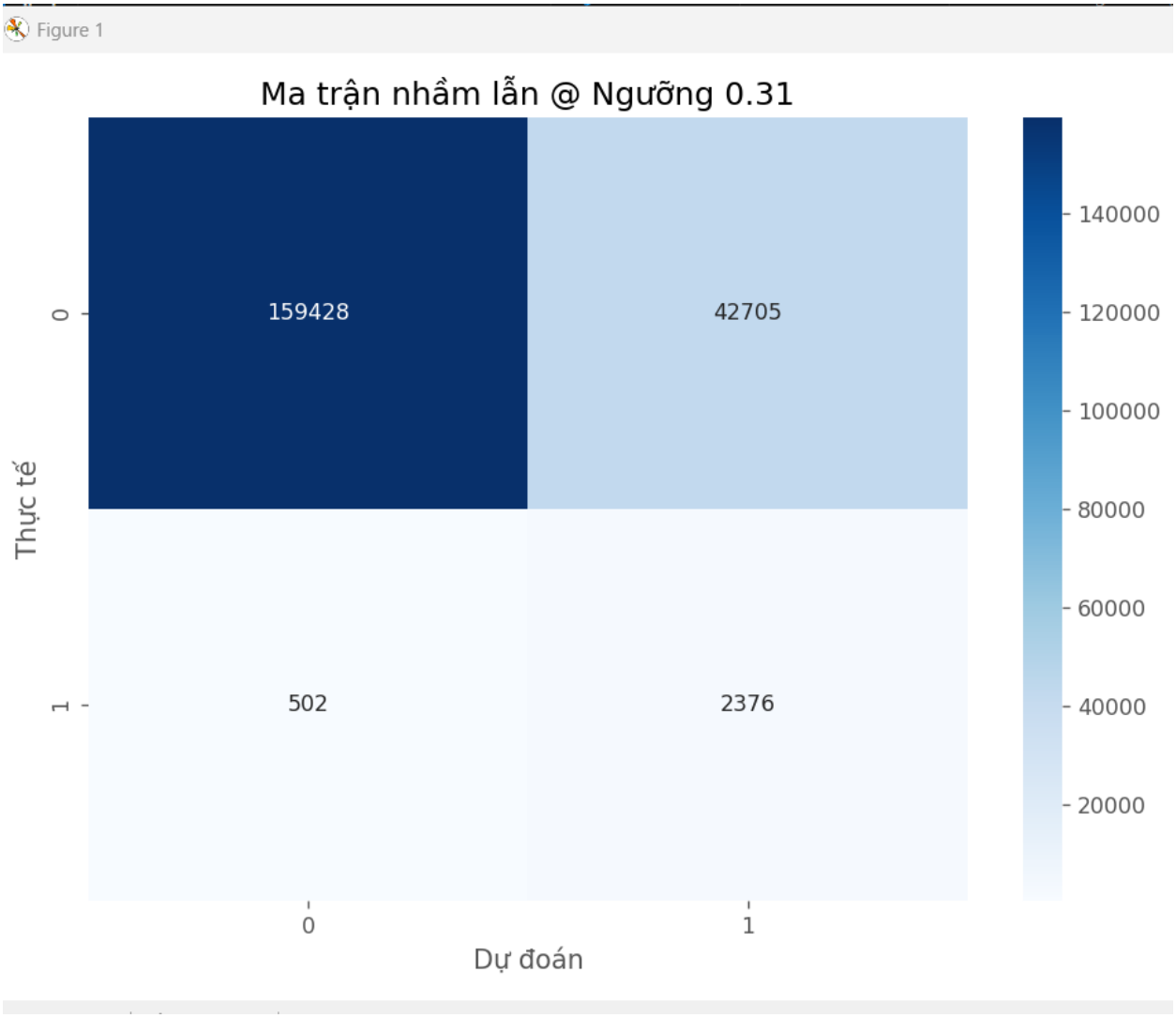
Mô hình phát hiện được **phần lớn các giao dịch gian lận** (recall cao), tuy nhiên **precision rất thấp**, cho thấy phần lớn các giao dịch bị dự đoán là gian lận thực tế lại không phải gian lận. Điều này phản ánh số lượng **cảnh báo giả (false positive) rất lớn**.

Đối với lớp Bình thường (0):

- Precision = 1,00
- Recall = 0,79

- F1-score = 0,88

Mô hình phân loại rất tốt các giao dịch bình thường, tuy nhiên vẫn còn một tỷ lệ đáng kể giao dịch hợp lệ bị gắn nhầm là gian lận.



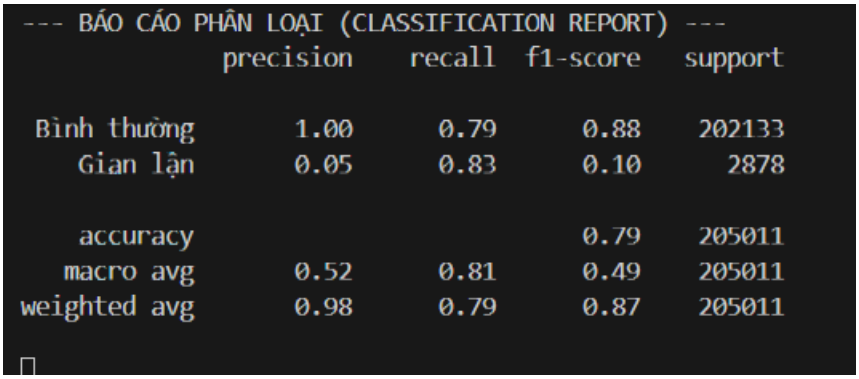
Trong đó:

- True Negative (TN) = 159.428: giao dịch bình thường được phân loại đúng
- False Positive (FP) = 42.705: giao dịch bình thường bị gắn nhầm là gian lận

- False Negative (FN) = 502: giao dịch gian lận bị bỏ sót
- True Positive (TP) = 2.376: giao dịch gian lận được phát hiện đúng

Kết luận : Mô hình **chưa phù hợp để tự động ra quyết định trong thực tế**, nhưng có thể sử dụng như một hệ thống **sàng lọc ban đầu** để chuyển các giao dịch nghi ngờ sang bước kiểm tra thủ công.

4.2 Mô hình XGBoost



```
--- BÁO CÁO PHÂN LOẠI (CLASSIFICATION REPORT) ---
              precision    recall  f1-score   support

 Bình thường       1.00      0.79      0.88    202133
    Gian lận       0.05      0.83      0.10      2878

    accuracy                           0.79    205011
   macro avg       0.52      0.81      0.49    205011
weighted avg       0.98      0.79      0.87    205011
```

Accuracy = 0,813 (81,3%)

Mặc dù độ chính xác tổng thể tương đối cao, nhưng do dữ liệu mất cân bằng nên chỉ số này chưa phản ánh đầy đủ hiệu quả của mô hình trong việc phát hiện gian lận.

Đối với lớp Fraud (1):

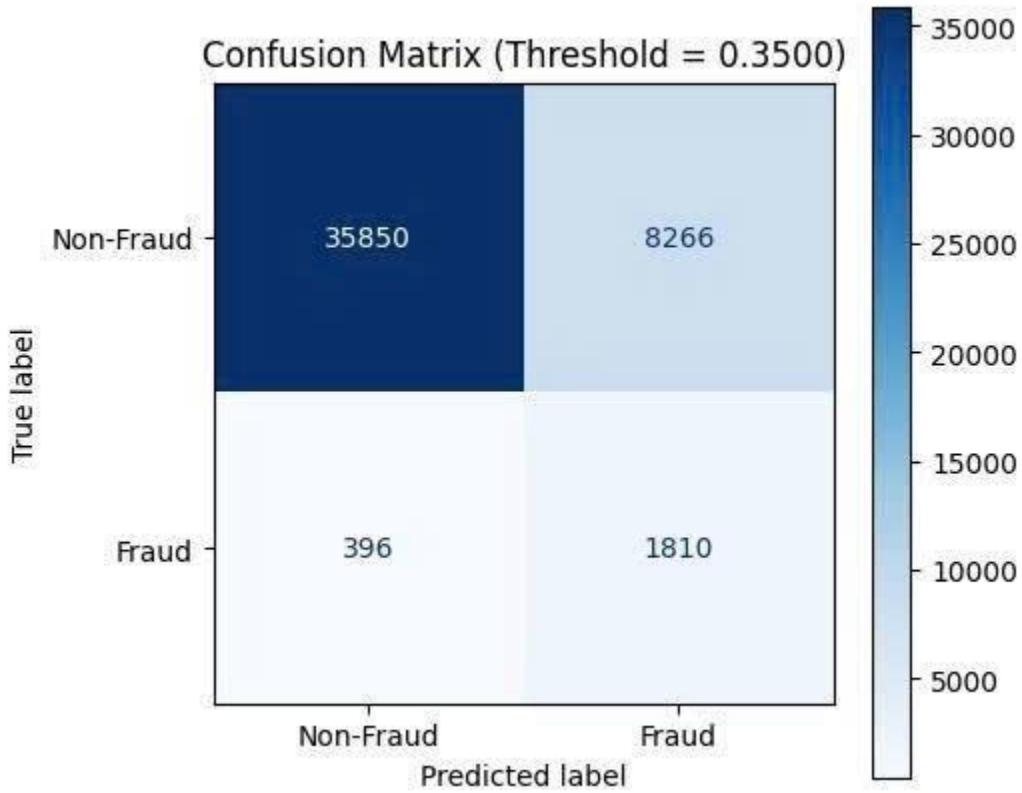
- Precision = 0,1796 (~18%)
- Recall = 0,8205 (~82%)
- F1-score = 0,2947

Mô hình có khả năng phát hiện phần lớn các giao dịch gian lận (recall cao), tuy nhiên tỷ lệ dự đoán đúng trong số các giao dịch bị gắn nhãn gian lận còn thấp (precision thấp), cho thấy tồn tại nhiều cảnh báo giả.

Đối với lớp Non-Fraud (0):

- Precision = 0,9891
- Recall = 0,8126

Mô hình phân loại tốt các giao dịch hợp lệ, nhưng vẫn còn một tỷ lệ đáng kể giao dịch hợp lệ bị gắn nhầm là gian lận.



Trong đó:

- True Positive (TP) = 1.810: giao dịch gian lận được phát hiện đúng
- False Negative (FN) = 396: giao dịch gian lận bị bỏ sót
- False Positive (FP) = 8.266: giao dịch hợp lệ bị gắn nhầm là gian lận
- True Negative (TN) = 35.850: giao dịch hợp lệ được phân loại đúng

Kết luận : Đây là mô hình cho ra kết quả tốt nhất trong ba mô hình, có thể dùng để phát triển

4.3 Catboost

```
=== BÁO CÁO PHÂN LOẠI ===
              precision    recall  f1-score   support

           0     0.9972    0.8460    0.9154    197794
           1     0.0540    0.7888    0.1012      2206

    accuracy                         0.8454    200000
   macro avg     0.5256    0.8174    0.5083    200000
weighted avg     0.9868    0.8454    0.9064    200000
```

Accuracy = 0,8454 (84,54%)
Mô hình đạt độ chính xác tổng thể khá cao. Tuy nhiên, với dữ liệu mất cân bằng nghiêm trọng, accuracy không đủ để đánh giá toàn diện khả năng phát hiện gian lận.

Đối với lớp Non-Fraud (0):

- Precision = 0,9972
- Recall = 0,8460
- F1-score = 0,9154

Mô hình phân loại rất tốt các giao dịch hợp lệ. Khi dự đoán là Non-Fraud thì gần như luôn đúng (precision rất cao). Tuy nhiên, vẫn còn khoảng **15,4% giao dịch hợp lệ** bị gắn nhầm là gian lận.

Đối với lớp Fraud (1):

- Precision = 0,0540 (~5,4%)
- Recall = 0,7888 (~78,9%)
- F1-score = 0,1012

Mô hình có khả năng phát hiện phần lớn các giao dịch gian lận (recall cao), nghĩa là ít bỏ sót gian lận. Tuy nhiên, precision rất thấp, cho thấy phần lớn các giao dịch bị dự đoán là gian lận thực tế lại không phải gian lận. Điều này phản ánh số lượng cảnh báo giả (false positive) rất lớn.

Trong đó:

- True Negative (TN) = 167.339: giao dịch hợp lệ được phân loại đúng
- False Positive (FP) = 30.455: giao dịch hợp lệ bị gắn nhầm là gian lận
- False Negative (FN) = 466: giao dịch gian lận bị bỏ sót
- True Positive (TP) = 1.740: giao dịch gian lận được phát hiện đúng

Kết luận: Mặc dù mô hình đạt độ chính xác tổng thể cao (84,54%) và có khả năng phát hiện phần lớn các giao dịch gian lận, nhưng hiệu quả thực tế đối với lớp Fraud còn hạn chế do precision quá thấp. Mô hình hiện tại phù hợp với kịch bản ưu tiên phát hiện gian lận hơn là độ chính xác của cảnh báo, tuy nhiên chưa đáp ứng tốt yêu cầu của một hệ thống tự động trong môi trường thực tế.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Trong khuôn khổ đề tài "Phát hiện gian lận trong mở tài khoản ngân hàng bằng kỹ thuật khai phá dữ liệu", nhóm đã tiến hành nghiên cứu và xây dựng hệ thống phát hiện gian lận dựa trên các thuật toán học máy có giám sát. Đề tài đã thực hiện đầy đủ các bước của quy trình khai phá dữ liệu, từ tiền xử lý, xây dựng mô hình đến đánh giá kết quả thực nghiệm.

Nhóm đã sử dụng bộ dữ liệu Bank Account Fraud Dataset (NeurIPS 2022) và tiến hành các bước tiền xử lý quan trọng như mã hóa biến phân loại, chuẩn hóa dữ liệu và xử lý mất cân bằng lớp. Trên cơ sở đó, các mô hình học máy hiện đại bao gồm LightGBM, XGBoost và CatBoost đã được huấn luyện và so sánh.

Kết quả thực nghiệm cho thấy các mô hình Gradient Boosting đều đạt hiệu năng tốt trong bài toán phát hiện gian lận. Đặc biệt, mô hình LightGBM cho thấy sự cân bằng hợp lý

giữa độ chính xác và tốc độ huấn luyện, đồng thời đạt kết quả tốt khi điều chỉnh ngưỡng quyết định dựa trên tỷ lệ báo động giả (False Positive Rate). Điều này chứng tỏ tính hiệu quả và khả năng ứng dụng thực tế của mô hình trong môi trường ngân hàng.

Thông qua đề tài, nhóm đã củng cố kiến thức về khai phá dữ liệu, hiểu rõ hơn về bài toán phân loại dữ liệu mất cân bằng, cũng như rèn luyện kỹ năng xây dựng và đánh giá các mô hình học máy trong lĩnh vực tài chính.

5.2. Hướng phát triển

Mặc dù đã đạt được những kết quả tích cực, đề tài vẫn còn một số hạn chế và có thể được phát triển thêm trong tương lai như sau:

- Mở rộng nghiên cứu với các kỹ thuật xử lý mất cân bằng nâng cao hơn như SMOTE, ADASYN hoặc kết hợp nhiều phương pháp sampling.
- Thử nghiệm các phương pháp tối ưu siêu tham số (Hyperparameter Tuning) tự động như Bayesian Optimization để cải thiện hiệu năng mô hình.
- Kết hợp thêm các mô hình học sâu (Deep Learning) hoặc mô hình lai (ensemble) để nâng cao khả năng phát hiện gian lận.
- Nghiên cứu ảnh hưởng của từng đặc trưng (feature importance) nhằm tăng khả năng giải thích mô hình.
- Triển khai mô hình dưới dạng ứng dụng web hoặc dashboard trực quan, hỗ trợ nhân viên ngân hàng theo dõi và đánh giá các hồ sơ có nguy cơ gian lận theo thời gian thực.
- Áp dụng mô hình cho các bài toán gian lận khác trong lĩnh vực tài chính như gian lận giao dịch, gian lận thẻ tín dụng.

CHƯƠNG 6: TÀI LIỆU THAM KHẢO

- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Ke, G., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Prokhorenkova, L., et al. (2018). CatBoost: Unbiased Boosting with Categorical Features. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Kaggle. *Bank Account Fraud Dataset (NeurIPS 2022)*.
- Giáo trình DATA MINING – Vũ Thị Hạnh
- Scikit-learn Documentation. https://scikit-learn.org/
- LightGBM Documentation. https://lightgbm.readthedocs.io/
- XGBoost Documentation. https://xgboost.readthedocs.io/
- CatBoost Documentation. https://catboost.ai/