**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**



**NGUYỄN VĂN HOÀI NAM**

**KHÓA LUẬN TỐT NGHIỆP**

**HỆ THỐNG PHÂN TÍCH XU HƯỚNG GIÁ TIỀN MÃ HÓA THỜI GIAN THỰC SỬ DỤNG APACHE KAFKA VÀ MACHINE LEARNING**

**Chuyên ngành: Khoa học dữ liệu**

**Giáo viên hướng dẫn: TS.Nguyễn Chí Kiên**

**TP.HỒ CHÍ MINH, THÁNG 05 NĂM 2025**

**INDUSTRIAL UNIVERSITY OF HO CHI MINH CITY FACULTY OF INFORMATION TECHNOLOGY**



**NGUYEN VAN HOAI NAM**

**GRADUATION THESIS**

**REAL-TIME CRYPTOCURRENCY TREND ANALYSIS SYSTEM USING APACHE KAFKA AND MACHINE LEARNING**

**Major: Data Science**

**Instructor: PhD. Nguyen Chi Kien**

**HO CHI MINH CITY, MAY 2025**

**CONTENT SUMMARY**

**Title**: Hệ thống phân tích xu hướng giá tiền mã hóa thời gian thực sử dụng Apache Kafka và Machine Learning

**Abstract**:

Trong bối cảnh thị trường tiền mã hóa (cryptocurrency) liên tục biến động và có khối lượng giao dịch lớn theo thời gian thực, nhu cầu xây dựng một hệ thống phân tích và dự đoán xu hướng giá trở nên ngày càng cấp thiết nhằm hỗ trợ nhà đầu tư đưa ra quyết định kịp thời và chính xác. Đề tài này đề xuất và xây dựng một hệ thống phân tích xu hướng giá tiền mã hóa thời gian thực, với dữ liệu được thu thập từ sàn giao dịch Binance theo chu kỳ 5 phút/giao dịch (5-minute candlestick).

Hệ thống bao gồm hai hướng chính:  
(1) **Phân tích và trực quan hóa dữ liệu thời gian thực**: Dữ liệu thị trường (giá mở cửa, đóng cửa, cao nhất, thấp nhất, khối lượng giao dịch) được thu thập liên tục thông qua Binance API, sau đó đẩy vào Apache Kafka để đảm bảo khả năng xử lý streaming với độ trễ thấp. Các dịch vụ tiêu thụ dữ liệu sử dụng PySpark để xử lý và tính toán các chỉ số kỹ thuật như Moving Average (MA), Relative Strength Index (RSI) và Bollinger Bands. Dữ liệu sau xử lý được lưu vào cơ sở dữ liệu NoSQL (Apache Cassandra) để phục vụ cho việc hiển thị biểu đồ giá thời gian thực trên dashboard.

(2) **Huấn luyện và ứng dụng mô hình học máy để dự đoán giá**: Dữ liệu lịch sử từ Kafka được luồng hóa vào PostgreSQL để lưu trữ ổn định và phục vụ cho việc phân tích offline. Apache Airflow được sử dụng để lập lịch thu thập, xử lý, và huấn luyện mô hình tự động theo chu kỳ hàng ngày. Các mô hình học máy được áp dụng bao gồm Linear Regression, Random Forest, và XGBoost nhằm dự đoán giá “close” tiếp theo dựa trên chuỗi dữ liệu 5 phút gần nhất. Các mô hình được huấn luyện với dữ liệu đầu vào gồm nhiều đặc trưng như giá, khối lượng, biến động giá, và các chỉ báo kỹ thuật.

Thông qua việc kết hợp các công nghệ hiện đại trong hệ sinh thái Big Data và Machine Learning, hệ thống không chỉ cung cấp cái nhìn trực quan về diễn biến thị trường tiền mã hóa mà còn hỗ trợ dự báo giá trong tương lai gần, góp phần tăng hiệu quả phân tích và ra quyết định đầu tư.

# LỜI CẢM ƠN

Lời đầu tiên em xin chân thành bày tỏ lòng biết ơn đến thầy **TS. Nguyễn Chí Kiên** người đã hết lòng giúp đỡ, hướng dẫn, truyền đạt kinh nghiệm, góp ý và tạo mọi điều kiện tốt nhất cho em hoàn thành khóa luận tốt nghiệp.

Em xin cảm ơn thầy **TS. Vũ Đức Thịnh** và thầy **TS. Nguyễn Hữu Vũ** đã đồng ý phản biện đề tài khóa luận tốt nghiệp của em. Em tin rằng những đánh giá phản biện của các thầy sẽ góp phần quan trọng trong việc hoàn thiện đề tài này.

Em xin cảm ơn đến toàn thể quý thầy cô trong khoa Công Nghệ Thông Tin – Trường Đại học Công nghiệp Thành phố Hồ Chí Minh đã tận tình truyền đạt những kiến thức quý báu cũng như tạo điều kiện thuận lợi cho em trong suốt quá trình học tập nghiên cứu và cho đến khi thực hiện khóa luận tốt nghiệp.

Mặc dù đã nỗ lực cố gắng cùng với sự tận tâm của thầy giáo hướng dẫn nhưng do trình độ còn hạn chế, nội dung đề tài còn khá mới mẻ với em nên khó tránh khỏi những sai sót trong quá trình tiếp nhận kiến thức. Em rất mong nhận được góp ý từ phía thầy cô để em có thể hoàn thiện đề tài.

Xin chân thành cảm ơn!

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

*TP.Hồ Chí Minh, Tháng 12 Năm 2024*

**GIÁO VIÊN HƯỚNG DẪN NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN**

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

.......................................................................................................................................

*TP.Hồ Chí Minh, Tháng 12 Năm 2024*

**GIÁO VIÊN PHẢN BIỆN**

# MỤC LỤC

[LỜI CẢM ƠN 3](#_heading=h.gjdgxs)

[MỤC LỤC 8](#_heading=h.30j0zll)

[MỤC LỤC HÌNH ẢNH 9](#_heading=h.1fob9te)

[DANH MỤC BẢNG BIỂU 13](#_heading=h.3znysh7)

[CHƯƠNG 1. GIỚI THIỆU 17](#_heading=h.2et92p0)

[1.1. Tổng quan 17](#_heading=h.tyjcwt)

[1.1.1. Bối cảnh 17](#_heading=h.3dy6vkm)

[1.1.2. Lý do chọn đề tài 18](#_heading=h.1t3h5sf)

[1.2. Mục tiêu nghiên cứu 19](#_heading=h.4d34og8)

[1.3. Phạm vi nghiên cứu 20](#_heading=h.2s8eyo1)

[1.4. Ý nghĩa khoa học và thực tiễn 20](#_heading=h.17dp8vu)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 22](#_heading=h.3rdcrjn)

[2.1. Chấm điểm tín dụng 22](#_heading=h.26in1rg)

[2.1.1. Điểm tín dụng 22](#_heading=h.lnxbz9)

[2.1.2. Chấm điểm tín dụng 23](#_heading=h.35nkun2)

[2.1.3. Nguyên cứu liên quan 24](#_heading=h.1ksv4uv)

[2.2. Spark và Hadoop 25](#_heading=h.ymzyduhw7stj)

[2.2.1. Spark 25](#_heading=h.lo09uv53k4wr)

[2.2.2. Hadoop 27](#_heading=h.muvnhg41zaeq)

[2.3. Học máy 28](#_heading=h.44sinio)

[2.4. Bài toán ứng dụng các thuật toán học máy vào mô hình chấm điểm tín dụng 30](#_heading=h.3j2qqm3)

[2.4.1. Tổng quan 30](#_heading=h.1y810tw)

[2.4.2. Nghiên cứu liên quan 32](#_heading=h.4i7ojhp)

[2.5. Thuật toán học máy được ứng dụng vào xây mô hình chấm điểm tín dụng 33](#_heading=h.2xcytpi)

[2.5.1. Logistic regression 33](#_heading=h.1ci93xb)

[2.5.2. Random forest 37](#_heading=h.3whwml4)

[2.5.3. XGBoost 41](#_heading=h.3as4poj)

[2.6. Kỹ thuật và phương pháp sử dụng trong nghiên cứu và thực nghiệm 44](#_heading=h.2p2csry)

[2.6.1. Feature engineering 44](#_heading=h.147n2zr)

[2.6.2. Gradient descent 46](#_heading=h.3o7alnk)

[2.6.3. Phương pháp đánh giá mô hình 48](#_heading=h.32hioqz)

[2.6.4. Phương pháp tìm ngưỡng phân loại 54](#_heading=h.1hmsyys)

[2.6.5. Phương pháp chuyển đổi điểm tín dụng từ kết quả đầu ra mô hình 55](#_heading=h.41mghml)

[CHƯƠNG 3. DỮ LIỆU 57](#_heading=h.2grqrue)

[3.1. Tổng quan dữ liệu 57](#_heading=h.vx1227)

[3.2. Mô tả dữ liệu 59](#_heading=h.3fwokq0)

[3.3. Chuẩn bị dữ liệu 62](#_heading=h.1v1yuxt)

[3.3.1. Phân tích khám phá dữ liệu 62](#_heading=h.4f1mdlm)

[3.3.2. Trích xuất đặc trưng dữ liệu 75](#_heading=h.2u6wntf)

[3.3.3. Tổng hợp dữ liệu 78](#_heading=h.19c6y18)

[CHƯƠNG 4. THỰC NGHIỆM VÀ KẾT QUẢ 78](#_heading=h.3tbugp1)

[4.1. Thực nghiệm 78](#_heading=h.28h4qwu)

[4.1.1. Dữ liệu 78](#_heading=h.nmf14n)

[4.1.2. Huấn luyện mô hình 79](#_heading=h.37m2jsg)

[4.2. Kết quả 85](#_heading=h.1mrcu09)

[4.2.1. Kết quả huấn luyện 85](#_heading=h.2lwamvv)

[4.2.2. Điểm tín dụng 88](#_heading=h.206ipza)

[4.2.3. Thuộc tính quan trọng 89](#_heading=h.4k668n3)

[4.3. Kết luận 91](#_heading=h.2zbgiuw)

[CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 92](#_heading=h.1egqt2p)

[5.1. Kết quả 92](#_heading=h.3ygebqi)

[5.1.1. Kết quả đạt được 92](#_heading=h.2dlolyb)

[5.2. Hướng phát triển 93](#_heading=h.sqyw64)

[TÀI LIỆU THAM KHẢO 93](#_heading=h.3cqmetx)

[NHẬT KÝ LÀM VIỆC 106](#_heading=h.1rvwp1q)

# MỤC LỤC HÌNH ẢNH

Hình 2.1 Biểu diễn đồ thị hàm Sigmoid .................................................................... 34

Hình 2.2 Biểu diễn đồ thị hàm Tanh ......................................................................... 34

Hình 2.3 Kiến trúc tổng quan về mô hình kết hợp .................................................... 38

Hình 2.4 Quá trình tạo thành một mô hình của thuật toán Random forest ............... 39

Hình 2.5 Huấn luyện mô hình XGBoost ................................................................. 41

Hình 2.6 Minh họa một bước giảm của loss function ............................................. 47

Hình 2.7 Confusion Matrix ..................................................................................... 49

Hình 2.8 Công thức tính hệ số Gini ...........................................................................

53

Hình 2.9 Biểu đồ đường cong Loenz..........................................................................

53

Hình 3.0 Công thức tính K-S Statistic........................................................................

54

Hình 3.1 Minh họa giá trị ngưỡng phân loại tốt nhất trên đường cong ROC ………

55

Hình 3.2 Logo Home Credit Group .......................................................................... 57

Hình 3.3 Tổng quan mối quan hệ giữa các tệp dữ liệu trong bộ dữ liệu Home Credit

Default Risk .............................................................................................................. 59

Hình 3.4 Đồ thị biểu diễn số lượng nhãn phân loại của 2 lớp trong biến mục tiêu .. 62

Hình 3.5 Biểu đồ số lượng loại khoản vay được thực hiện (phải) và tỷ lệ của các loại

khoản vay trên các trường hợp không thể trả nợ (trái) ............................................. 63 Hình 3.6 Biểu đồ số lượng khách hàng sở hữu ô tô (phải) và tỷ lệ trên các khoản vay

không thể trả (trái) ..................................................................................................... 64

Hình 3.7 Biểu đồ số lượng khách hàng sở hữu bất động sản (phải) và tỷ lệ trên các

khoản vay không thể trả (trái) ................................................................................... 64

Hình 3.8 Biểu đồ tình trạng gia đình của các khách hàng vay (phải) và tỷ lệ trên các khoản vay không thể thanh toán (trái) ...................................................................... 65

Hình 3.9 Biểu đồ phân phối số lượng thành viên trong gia đình (phải) của các khách hàng vay và tỷ lệ trên các khoản vay không thể thanh toán (trái) ............................ 66 Hình 3.10 Biểu đồ thể hiện số lượng khách hàng vay với các loại thu nhập khác nhau

và tỷ lệ trên các khoản vay không hoàn trả ............................................................... 67

Hình 3.11 Biểu đồ phân phối ngành nghề làm việc của khách hàng vay và tỷ lệ trên số lượng không hoàn trả nợ ....................................................................................... 68

Hình 3.12 Biểu đồ phân phối số lượng của các loại nhà ở/cư trú của các khách hàng vay và tỷ lệ trên số lượng không hoàn trả nợ ............................................................ 69 Hình 3.13 Phân phối loại tín dụng và tỷ lệ trên các khoản vay không hoàn trả nợ .. 70 Hình 3.14 Biểu đồ phân phối của các loại tín dụng khác nhau và tỷ lệ trên các khoản vay không hoàn trả nợ ............................................................................................... 71

Hình 3.15 Phân phối thời hạn tín dụng của các khoản tín dụng đã vay trước đây từ dữ liệu của phòng tín dụng ............................................................................................. 72

Hình 3.16 Phân bố số lượng của các loại hợp đồng tín dụng khác nhau và tỷ lệ trên

các khoản vay không hoàn trả nợ .............................................................................. 73 Hình 3.17 Phân bố số lượng của các mục đích vay tiền mặt khác nhau và tỷ lệ trên số

lượng khoản vay không hoàn trả nợ .......................................................................... 74

Hình 4.1 Đồ thị biểu diễn đường cong ROC của các mô hình huấn luyện trên toàn bộ

dữ liệu ...................................................................................................................... 86

# DANH MỤC BẢNG BIỂU

Bảng 2.1 Bảng phân chia giá trị thông tin của các thuộc tính dựa trên khoảng giá trị

của chỉ số IV .............................................................................................................. 46

Bảng 3.1 Thông tin về các tập dữ liệu trong bộ dữ liệu Home Credit Default

Risk............................................................................................................................ 60

Bảng 4.1 Tham số mô hình Logistic regression ........................................................ 80

Bảng 4.2 Tham số mô hình Random forest .............................................................. 82

Bảng 4.3 Tham số huấn luyện mô hình XGBoost .................................................... 84

Bảng 4.4 Chỉ số đánh giá trên biến phân loại ......................................................... 86

Bảng 4.5 Kết quả tổng hợp chỉ số đánh giá mô hình .............................................. 87

Bảng 4.6 Giá trị dự đoán của mô hình XGBoost được chuyển đổi thành điểm tín dụng của 10 mẫu dữ liệu với ngưỡng phân loại 0.6 ................................................ 89 Bảng 4.7 Các thuộc tính quan trọng của mô hình ................................................. 91

**DANH MỤC THUẬT NGỮ VÀ VIẾT TẮT**

|  |  |
| --- | --- |
| **TỪ NGỮ** | **Ý NGHĨA** |
| Activation function | Hàm kích hoạt |
| Bin | Ngăn chứa dữ liệu |
| Binary tree | Cây nhị phân |
| Bootstrap | Lấy mẫu tái lập |
| Categorical feature | Thuộc tính phân loại |
| Concatenate layer | Lớp nối của mạng neural network truyền thẳng |
| Decision trees | Thuật toán cây quyết định |
| Ensemble learning | Phương pháp học đồng bộ |
| Feed-forward neural network | Mạng neural network truyền thẳng |
| FN | False negative |
| FP | False positive |
| FPR | False positive rate |
| GAN - generative adversarial networks | Mạng đối nghịch |
| GBDT | Gradient boosting decision tree - Cây quyết định tăng  cường độ dốc |

|  |  |
| --- | --- |
| GD - Gradient descent | Thuật toán tối ưu hàm mất mát được sử dụng trong các bài toán Machine Learning và Deep Learning |
| GOSS | Gradient-based one-side sampling - Lấy mẫu một  phía dựa trên độ dốc |
| Hidden layers | Lớp ẩn của neural network |
| Input layer | Lớp đầu vào của neural network |
| Iteration | Vòng lặp |
| IV | Information value - Chỉ số giá trị thông tin |
| Label | Nhãn |
| Loss function | Hàm mất mát |
| LR - Learning rate | Tốc độ học |
| Max-depth | Độ sâu lớn nhất |
| NN | Neural network |
| Node | Nốt |
| Output layer | Lớp đầu ra của neural network |
| Overfitting - Trang bị quá  mức | Là hiện tượng mô hình tìm được quá khớp với dữ liệu training |
| RNN - recurrent neural networks | Mạng nơron hồi quy |
| SGD - Stochastic Gradient Descent | Thuật toán giảm độ dốc ngẫu nhiên |
| std | Standard deviation - Độ lệch chuẩn |
| Test set | Tập dữ liệu kiểm thử |
| Threshold | Ngưỡng phân loại |
| TN | True negative |
| TP | True positive |
| TPR | True positive rate |
| Train set | Tập dữ liệu huấn luyện |
| WOE | Weight of evidence - Trọng số dấu hiệu |

# CHƯƠNG 1. GIỚI THIỆU

## 1.1. Tổng quan

### 1.1.1. Bối cảnh

Tín dụng đóng một vai trò quan trọng trong hệ thống tài chính. Hiện nay, tín dụng đang tăng trưởng cực kỳ mạnh mẽ với sự gia tăng liên tục về số lượng khách hàng đi và nhu cầu tín dụng. Xu hướng này không chỉ giới hạn ở các nước phát triển mà còn được thấy ở các thị trường mới nổi. Tại Việt Nam, số lượng người sử dụng tín dụng đã tăng nhanh trong những năm gần đây, nhờ tăng trưởng kinh tế của đất nước và sự gia tăng của tầng lớp trung lưu. Theo Ngân hàng Nhà nước Việt Nam, tính đến cuối năm 2020, dư nợ tín dụng đạt 8.000 nghìn tỷ đồng, tăng 10,1% so với cuối năm 2019 [1]. Sự tăng trưởng này được thúc đẩy bởi một số yếu tố, bao gồm sự cạnh tranh ngày càng tăng giữa những tổ chức tín dụng, sự gia tăng của các nền tảng cho vay trực tuyến và sự phát triển của các sản phẩm và dịch vụ tín dụng mới. Các tổ chức cho vay hiện đang cung cấp nhiều loại sản phẩm tín dụng hơn để đáp ứng các nhu cầu khác nhau của khách hàng, chẳng hạn như cho vay cá nhân, cho vay mua ô tô, cho vay mua nhà… Ngoài ra, việc áp dụng các dịch vụ cho vay dựa trên các nền tảng số đã giúp khách hàng tiếp cận tín dụng dễ dàng hơn, giờ đây khách hàng có thể đăng ký khoản vay trực tuyến và nhận tiền trong vòng vài ngày.

Đi kèm với sự phát triển mạnh của các hoạt động tín dụng đó là vấn đề kiểm soát rủi ro làm lan rộng và tình trạng bất ổn kinh tế. Những tổ chức hoặc cá nhân cho vay đang đối mặt với rủi ro vỡ nợ gia tăng.

Chấm điểm tín dụng là một giai đoạn quan trọng trong quy trình quản lý rủi ro của các ngân hàng, các định chế tài chính, các tổ chức tín dụng. Chấm điểm tín dụng tốt sẽ góp phần làm cho chất lượng cho vay tốt hơn. Chất lượng cho vay là yếu tố quyết định hàng đầu đến sự cạnh tranh, tồn tại và lợi nhuận của các ngân hàng, các định chế tài chính, các tổ chức tín dụng [2].

Các mô hình chấm điểm tín dụng được ra đời nhằm xác định mức độ tin cậy về tín dụng của các cá nhân cũng như doanh nghiệp từ đó tối thiểu hóa rủi ro khoản vay. Theo Muhammad Azeem Qureshi [3], chấm điểm tín dụng là một công cụ quan trọng được các ngân hàng và tổ chức tài chính sử dụng để đánh giá mức độ tin cậy của người đi vay. Nó liên quan đến việc sử dụng các mô hình thống kê để phân tích lịch sử tín dụng, hành vi thanh toán và các dữ liệu tài chính khác của người vay để tạo điểm tín dụng. Mục đích của việc chấm điểm tín dụng là giảm rủi ro vỡ nợ và tối đa hóa lợi nhuận cho tổ chức tín dụng. Ví dụ, điểm tín dụng FICO dựa trên các yếu tố lịch sử thanh toán, các khoản nợ, độ dài lịch sử tín dụng, cơ cấu tín dụng, tín dụng mới để đánh giá điểm tín dụng của người đi vay [4]. Hiện nay, với với những tiến bộ công nghệ, các mô hình chấm điểm tín dụng đã trở nên tự động hơn và dựa trên các nguồn dữ liệu thay thế. Các thuật toán học máy ngày càng được sử dụng nhiều hơn trong các mô hình chấm điểm tín dụng nhằm phân tích lượng dữ liệu khổng lồ và xác định các mẫu dữ liệu không rõ ràng. Việc sử dụng các thuật toán học máy (Machine learning - ML) có khả năng cải thiện đáng kể độ chính xác của các mô hình chấm điểm tín dụng. Điều này sẽ mở rộng khả năng tiếp cận tín dụng cho những cá nhân có thể không có thông tin về lịch sử tín dụng [5].

### 1.1.2. Lý do chọn đề tài

Điểm tín dụng đóng một vai trò quan trọng trong việc xác định mức độ tin cậy của người đi vay và là điều cần thiết để các ngân hàng, các định chế tài chính, các tổ chức tín dụng đưa ra các quyết định cho vay đúng đắn.

Các mô hình chấm điểm tín dụng truyền thống thường dựa vào các thuộc tính cố định và hạn chế như lịch sử tín dụng, hành vi thanh toán và các dữ liệu tài chính khác để tạo điểm tín dụng. Cách tiếp cận này có thể dẫn đến những dự đoán không chính xác và đưa ra đánh giá không công bằng về mức độ tín nhiệm, vì nó không nắm bắt được bản chất phức tạp tín dụng của từng cá nhân. Hơn nữa, các mô hình chấm điểm tín dụng truyền thống không tính đến các yếu tố phi tài chính như lịch sử việc làm, giáo dục và các chỉ số kinh tế xã hội khác, điều này có thể dẫn đến những quyết định tín dụng rủi ro hoặc làm giảm khả năng tiếp cận tín dụng của một cá nhân. Ngoài ra, các mô hình chấm điểm tín dụng truyền thống dễ bị chi phối bởi các yếu tố liên quan đến giá trị đạo đức. Theo Hội đồng Quan hệ Đối ngoại (CFR), Ba cơ quan xếp hạng tín dụng là Moody’s Investor Services, Standard and Poor’s (S&P), Fitch Group đã bị cáo buộc góp phần chính gây ra cuộc khủng hoảng tài chính toàn cầu năm 2008 khi các cơ quan này đã xếp hạng tín dụng sai lệch cho các khoản tín dụng [6]. Các nhược điểm của mô hình chấm điểm tín dụng truyền thống có thể gây những ảnh hưởng rất nặng nề đến hệ thống tài chính.

Các thuật toán học máy có thể được áp dụng để giải quyết những vấn đề này. Các thuật toán học máy có thể phân tích với khối lượng dữ liệu lớn, bao gồm cả những yếu tố phi tài chính để tạo ra điểm tín dụng chính xác và mang tính dự đoán hơn [7]. Các thuật toán này cũng có thể đưa ra những thuộc tính quan trọng của mô hình từ đó cung cấp cái nhìn sâu sắc hơn về các yếu tố ảnh hưởng đến quyết định tín dụng. Chính vì thấy được những ưu điểm của phương pháp này, em chọn nghiên cứu “Xây dựng mô hình chấm điểm tín dụng sử dụng các thuật toán học máy bằng spark” làm đề tài khóa luận tốt nghiệp của mình.

## 1.2. Mục tiêu nghiên cứu

* Hiểu về các kiến thức tài chính liên quan đến hoạt động cho vay tín dụng.
* Nắm rõ được kiến thức, cách hoạt động và ứng dụng của spark và hadoop.
* Nắm rõ được kiến thức, cách hoạt động và ứng dụng của các thuật toán học máy:
  + Logistic regression
  + Random forest
  + XGBoost
* Phân tích và khám phá dữ liệu, thu thập thông tin chi tiết từ quá trình phân tích khám phá.
* Trích xuất và đánh giá được các đặc trưng từ dữ liệu.
* Xây dựng và huấn luyện mô hình dữ liệu từ các thuật toán học máy đã tìm hiểu.
* Kết hợp mô hình đã huấn luyện bằng các phương pháp học đồng bộ.
* So sánh hiệu suất giữa các mô hình đã đào tạo, đưa ra đánh giá.
* Chấm điểm tín dụng sử dụng dữ liệu có sẵn từ mô hình đã huấn luyện có hiệu suất tốt nhất.
* Đưa ra các thuộc tính quan trọng trong dữ liệu.

## 1.3. Phạm vi nghiên cứu

* Nghiên cứu đào tạo mô hình dữ liệu sử dụng các thuật toán:
  + Logistic regression
  + Random forest
  + XGBoost
* Sử dụng bộ dữ liệu “Home Credit Default Risk” có sẵn và được công khai trên Kaggle [8].

## 1.4. Ý nghĩa khoa học và thực tiễn

Từ góc độ khoa học, việc phát triển các mô hình chấm điểm tín dụng chính xác và đáng tin cậy là một lĩnh vực nghiên cứu tích cực trong lĩnh vực học máy và khoa học dữ liệu. Các thuật toán học máy ngày càng được sử dụng để phát triển các mô hình chấm điểm tín dụng chính xác hơn khi có thể kết hợp được nhiều nguồn dữ liệu và thuộc tính hơn. Đề tài này có khả năng đóng góp vào việc phát triển các thuật toán, kỹ thuật và phương pháp mới để các mô hình đánh giá rủi ro tín dụng ngày càng chính xác, công bằng và minh bạch hơn.

Từ góc độ thực tế, các mô hình chấm điểm tín dụng được các tổ chức tín dụng và các tổ chức tài chính sử dụng để đánh giá mức độ tin cậy của người vay và đưa ra các quyết định cho vay đúng đắn. Sử dụng các mô hình chấm điểm tín dụng chính xác có thể giúp người cho vay xác định rủi ro tín dụng của những người đăng ký khoản vay. Việc sử dụng các thuật toán học máy trong chấm điểm tín dụng cũng có thể giúp người cho vay tự động hóa quy trình đăng ký khoản vay, giảm chi phí và nâng cao hiệu quả tổng thể của hoạt động cho vay của họ.

Khi kết hợp với các mô hình chấm điểm tín dụng truyền thống, việc sử dụng các thuật toán học máy có thể nâng cao độ chính xác và độ tin cậy của các đánh giá rủi ro tín dụng. Bằng cách kết hợp với các mô hình chấm điểm tín dụng truyền thống có thể cung cấp bức tranh toàn diện hơn về mức độ tin cậy của người đăng ký khoản vay, từ đó có thể xác định được các mẫu và mối quan hệ dữ liệu mà các mô hình chấm điểm tín dụng truyền thống có thể đã bỏ qua. Điều này có thể giúp các quyết định cho vay đúng hơn và giảm thiểu rủi ro tín dụng.

# CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

## 2.1. Chấm điểm tín dụng

### 2.1.1. Điểm tín dụng

Điểm tín dụng là một thang điểm số từ 300 đến 850 phản ánh mức độ đáng tin cậy của cá nhân hoặc tổ chức vay nợ. Điểm tín dụng ra đời vào năm 1989 khi tập đoàn Fair Isaac (nay là FICO) đã dựa trên thông tin từ các báo cáo tín dụng của người tiêu dùng để tạo ra khung điểm số đánh giá khả năng trả nợ của một cá nhân. Từ điểm tín dụng, bên cho vay có thể xác định được mức độ rủi ro mà bên vay nợ có thể có, từ đó xác định các yếu tố cho vay và xác định bên vay nợ sẽ trả lại khoản vay đúng hạn hay không [9].

Theo nghiên cứu do Cục Bảo vệ Tài chính Người tiêu dùng (CFPB) thực hiện, điểm tín dụng có tính dự báo cao về rủi ro tín dụng [10]. Nghiên cứu cho thấy những người vay có điểm tín dụng cao hơn ít có khả năng vỡ nợ hơn đáng kể, trong khi những người vay có điểm tín dụng thấp hơn có nhiều khả năng vỡ nợ hơn đáng kể. Ngoài ra, nghiên cứu cho thấy rằng điểm tín dụng có khả năng dự đoán rủi ro tín dụng cao hơn các yếu tố khác, chẳng hạn như thu nhập hoặc lịch sử việc làm.

Điểm tín dụng cũng có thể có tác động lớn đến tình trạng tài chính của một cá nhân. Nghiên cứu được thực hiện bởi Ngân hàng Dự trữ Liên bang Philadelphia [11] cho thấy những cá nhân có điểm tín dụng cao hơn sẽ có khả năng được chấp thuận cho các khoản vay và thẻ tín dụng hơn, đồng thời được cung cấp mức lãi suất thấp hơn và các điều khoản tốt hơn. Mặt khác, những cá nhân có điểm tín dụng thấp hơn có nhiều khả năng bị từ chối tín dụng hoặc đưa ra mức lãi suất cao hơn và các điều khoản kém thuận lợi.

### 2.1.2. Chấm điểm tín dụng

#### 2.1.2.1. Khái niệm

Chấm điểm tín dụng là một phương pháp thống kê được sử dụng để đánh giá mức độ tín nhiệm của một cá nhân hoặc doanh nghiệp liên quan đến các hoạt động tín dụng. Hand & Jacka là hai nhà kinh tế học và thống kê kê tài chính đã tuyên bố rằng: “quá trình (của các tổ chức tài chính) lập mô hình mức độ tin cậy được gọi là chấm điểm tín dụng” [12].

Chấm điểm tín dụng liên quan đến việc phân tích lịch sử tín dụng, thu nhập, tỷ lệ nợ trên thu nhập và các dữ liệu tài chính khác của người đi vay để dự đoán khả năng trả nợ đúng hạn của họ. Kết quả của quá trình chấm điểm tín dụng là điểm tín dụng nằm trong khoảng từ 300 đến 850 ở Hoa Kỳ.

Chấm điểm tín dụng là một công cụ quan trọng đối với người cho vay và các tổ chức tài chính vì sẽ giúp đưa ra quyết định sáng suốt về việc cho vay, quyết định hạn mức và lãi suất cho vay.

#### 2.1.2.2. Các mô hình chấm điểm tín dụng

Hiện nay, có nhiều mô hình tính điểm tín dụng khác nhau được sử dụng, nhưng mô hình phổ biến nhất là điểm FICO. Điểm FICO do Fair Isaac Corporation phát triển. Mô hình tính điểm này được sử dụng bởi đại đa số những tổ chức tín dụng ở Hoa Kỳ [13].

Điểm FICO nằm trong khoảng từ 300 đến 850, điểm số càng cao cho thấy rủi ro tín dụng càng thấp. Điểm FICO từ 740 trở lên thường được coi là rất tốt, trong khi điểm dưới 580 được coi là kém [14]. Mô hình chấm điểm này dựa trên nhiều

thông tin khác nhau đối với từng cá nhân. Tuy nhiên, để đánh giá điểm tín dụng, một số yếu tố chung được đặt ra với trọng số tương ứng như lịch sử thanh toán (35%), số tiền nợ (30%), độ dài lịch sử tín dụng (15%), kết hợp tín dụng (10%), tín dụng mới (10%).

Điểm FICO được cập nhật thường xuyên để phản ánh những thay đổi trong lịch sử tín dụng của người vay. Ví dụ: nếu người vay trả hết số dư thẻ tín dụng, điểm FICO của họ có thể tăng lên. Tương tự, nếu người vay bỏ lỡ khoản thanh toán hoặc nhận khoản nợ mới, điểm FICO của họ có thể giảm.

VantageScore cũng là mô hình chấm điểm tín dụng phổ biến hiện nay ngoài điểm FICO. Mô hình VantageScore được giới thiệu vào năm 2006 bởi ba văn phòng tín dụng chính là Equifax, Experian và TransUnion [15]. Giống như điểm FICO,

VantageScore được các tổ chức tín dụng sử dụng để đánh giá mức độ tin cậy của người đi vay và đưa ra quyết định cho vay sáng suốt. Các yếu tố đá giá chính đi kèm với trọng số của mô hình này bao gồm lịch sử thanh toán (41%), tuổi và loại tín dụng (20%), tỷ lệ hạn mức tín dụng sử dụng (20%), tổng số dư (6%), hành vi tín dụng gần đây (11%), tín dụng khả dụng (2%) [16].

Ngoài 2 mô hình chấm điểm tín dụng này còn có nhiều mô hình khác tuy nhiên đây là những mô hình được sử dụng rộng rãi trong hoạt động tín dụng và được đa phần các tổ chức tín dụng sử dụng.

### 2.1.3. Nguyên cứu liên quan

Nghiên cứu của Zhu và cộng sự [17] đã khám phá tác động của các nguồn dữ liệu thay thế đối với các mô hình chấm điểm tín dụng. Các tác giả nhận thấy rằng việc kết hợp dữ liệu thay thế, chẳng hạn như thanh toán tiền thuê nhà và hóa đơn tiện ích, đã cải thiện độ chính xác của các mô hình chấm điểm tín dụng và có thể giúp mở rộng khả năng tiếp cận tín dụng cho những người dân chưa được phục vụ đầy đủ. Các tác giả gợi ý rằng việc sử dụng các nguồn dữ liệu thay thế có thể là một công cụ quan trọng để giảm sai lệch trong các mô hình chấm điểm tín dụng.

Trong một nghiên cứu của Lin và các cộng sự [18] đã điều tra tác động của dữ liệu truyền thông xã hội đối với các mô hình chấm điểm tín dụng. Các tác giả nhận thấy rằng việc kết hợp dữ liệu truyền thông xã hội (chẳng hạn như các hoạt động trên nền tảng Facebook) vào các mô hình chấm điểm tín dụng đã cải thiện độ chính xác của mô hình trong việc dự đoán rủi ro tín dụng. Các tác giả gợi ý rằng dữ liệu truyền thông xã hội có thể là một công cụ hữu ích để giảm sự sai lệch trong các mô hình chấm điểm tín dụng và mở rộng khả năng tiếp cận tín dụng cho những người chưa có hoặc rất ít lịch sử tín dụng.

Một nghiên cứu của Bhattacharya và Bose [19] đã điều tra tác động của các yếu tố kinh tế vĩ mô đến rủi ro tín dụng. Các tác giả nhận thấy rằng các yếu tố kinh tế vĩ mô, chẳng hạn như tăng trưởng GDP và lạm phát, là những yếu tố dự báo đáng kể về rủi ro tín dụng. Các tác giả gợi ý rằng việc kết hợp các yếu tố kinh tế vĩ mô vào các mô hình chấm điểm tín dụng có thể cải thiện độ chính xác của chúng và giúp các cá nhân, tổ chức cho vay đưa ra quyết định cho vay đúng đắn hơn.

## 2.2. Spark và Hadoop

### 2.2.1. Spark

Apache Spark là một framework xử lý dữ liệu mã nguồn mở, được phát triển nhằm cung cấp khả năng xử lý dữ liệu nhanh, linh hoạt và dễ sử dụng. Spark nổi bật nhờ khả năng xử lý dữ liệu trong bộ nhớ (in-memory processing), giúp tăng tốc độ xử lý dữ liệu lớn lên đáng kể so với mô hình truyền thống dựa trên đĩa như Hadoop [20].

Spark có thể tích hợp tốt với nhiều nguồn dữ liệu như Hadoop Distributed File System (HDFS), Apache Cassandra, Amazon S3, và nhiều hệ thống lưu trữ khác, đồng thời hỗ trợ nhiều ngôn ngữ lập trình như Python, Scala, Java, và R [21].

Các thành phần chính:

- Spark Core là thành phần cốt lõi của Spark, chịu trách nhiệm quản lý tài nguyên, lập lịch công việc, và xử lý dữ liệu phân tán. Spark Core cung cấp API để thao tác với RDD (Resilient Distributed Dataset) – cấu trúc dữ liệu giúp lưu trữ và xử lý dữ liệu lớn theo cách phân tán [22].

- Spark SQL cung cấp một giao diện xử lý dữ liệu dạng cấu trúc, cho phép truy vấn dữ liệu bằng ngôn ngữ SQL hoặc các API dạng bảng (DataFrame và Dataset). Spark SQL hỗ trợ tích hợp với các nguồn dữ liệu như Hive, JDBC, và các tệp JSON hoặc Parquet [20].

- Spark Streaming:Thành phần này hỗ trợ xử lý dữ liệu thời gian thực, giúp Spark phân tích luồng dữ liệu liên tục từ Kafka, Flume, hoặc các nguồn dữ liệu thời gian thực khác [21].

- MLlib là thư viện máy học tích hợp của Spark, cung cấp các thuật toán phổ biến như phân cụm (clustering), phân loại (classification), hồi quy (regression), và nhiều công cụ tiện ích cho xử lý dữ liệu trước khi học máy [22].

- GraphX là một API chuyên xử lý dữ liệu dạng đồ thị (graph) và các tính toán liên quan như PageRank, Connected Components, và Triangles Counting [22].

Tính năng nổi bật:

- Xử lý dữ liệu nhanh: Nhờ khả năng xử lý trong bộ nhớ, Spark nhanh hơn Hadoop MapReduce từ 10–100 lần trong nhiều trường hợp [20].

- Hỗ trợ đa ngôn ngữ: Spark hỗ trợ nhiều ngôn ngữ lập trình phổ biến như Python, Scala, Java, và R [21].

- Khả năng tích hợp linh hoạt: Spark tích hợp dễ dàng với nhiều nguồn dữ liệu lớn như HDFS, Cassandra, và các cơ sở dữ liệu SQL [20].

- Hỗ trợ xử lý đa dạng: Spark xử lý dữ liệu hàng loạt, dữ liệu thời gian thực, máy học, và đồ thị trên cùng một nền tảng [21].

- Khả năng chịu lỗi cao: Nhờ sử dụng RDD, Spark có thể khôi phục dữ liệu tự động khi gặp sự cố [22].

Ứng dụng:

- Xử lý log dữ liệu và phân tích hành vi người dùng trong các nền tảng web [21].

- Phân tích thời gian thực dữ liệu tài chính, y tế, viễn thông [22].

- Phân tích dự đoán và học máy trong các lĩnh vực như thương mại điện tử và quảng cáo trực tuyến [22].

### 2.2.2. Hadoop

Hadoop là một framework mã nguồn mở của Apache, được thiết kế để xử lý và lưu trữ dữ liệu lớn trên các hệ thống phân tán. Hadoop giúp tối ưu hóa hiệu suất và giảm thiểu chi phí bằng cách sử dụng phần cứng thông thường và phân phối công việc xử lý trên nhiều nút trong một cụm máy tính. Hadoop có thể mở rộng từ một máy chủ đơn lẻ lên hàng ngàn máy chủ mà không ảnh hưởng đến hiệu suất.

Các thành phần chính:

- HDFS (Hadoop Distributed File System):HDFS là hệ thống tệp phân tán của Hadoop, nơi dữ liệu được chia thành các khối và phân phối qua các nút khác nhau trong cụm. HDFS có khả năng chịu lỗi rất tốt, với cơ chế sao chép dữ liệu giữa các nút để đảm bảo dữ liệu không bị mất khi xảy ra sự cố phần cứng[23].

- YARN (Yet Another Resource Negotiator):YARN là thành phần chịu trách nhiệm quản lý tài nguyên của các ứng dụng trong Hadoop, giúp điều phối công việc giữa các máy chủ và các ứng dụng phân tán. YARN cho phép Hadoop chạy nhiều loại ứng dụng khác nhau trên cùng một cụm máy tính[24].

- MapReduce:Là mô hình lập trình cho phép xử lý và phân tích dữ liệu lớn theo các bước *Map* (phân tách và xử lý dữ liệu) và *Reduce* (tổng hợp và trả kết quả). MapReduce giúp tối ưu hóa hiệu quả xử lý và hỗ trợ phân tích dữ liệu song song trên nhiều máy tính[25].

- Hadoop Common:Bao gồm các thư viện và công cụ cần thiết cho các thành phần khác của Hadoop, giúp dễ dàng cài đặt và vận hành hệ thống Hadoop[26].

Tính năng nổi bật:

- Khả năng mở rộng: Hadoop có thể mở rộng một cách linh hoạt, cho phép thêm các nút vào cụm để tăng cường khả năng xử lý khi cần thiết.

- Khả năng chịu lỗi: HDFS sao chép dữ liệu trên nhiều nút, giúp hệ thống hoạt động ổn định ngay cả khi gặp sự cố phần cứng.

- Chi phí thấp: Với Hadoop, bạn có thể sử dụng phần cứng thông thường thay vì phần cứng đắt tiền, giúp giảm thiểu chi phí cho các tổ chức.

- Xử lý dữ liệu gần nơi lưu trữ: Thay vì truyền tải dữ liệu về một nơi duy nhất, Hadoop thực hiện xử lý ngay trên các nút nơi dữ liệu được lưu trữ, giúp giảm thiểu độ trễ[27].

Ứng dụng:

- Phân tích log dữ liệu trong các dịch vụ web, mạng xã hội.

- Quá trình xử lý dữ liệu trong các hệ thống dữ liệu lớn và kho dữ liệu.

- Ứng dụng trong phân tích dự đoán và học máy để xử lý lượng dữ liệu khổng lồ trong các ngành tài chính, y tế, và viễn thông[23][25].

## 2.3. Học máy

Học máy là một lĩnh vực khoa học máy tính đang phát triển nhanh chóng, tập trung vào việc phát triển các thuật toán và mô hình cho phép máy tính học hỏi từ dữ liệu và đưa ra dự đoán hoặc quyết định. Theo IBM [28], học máy là một nhánh của trí tuệ nhân tạo (AI) và khoa học máy tính tập trung vào việc sử dụng dữ liệu và thuật toán để bắt chước cách con người học, dần dần cải thiện độ chính xác của nó.

Có thể chia hệ thống của các thuật toán học máy thành 3 thành phần chính:

* Quy trình quyết định (Decision process): Các thuật toán dựa trên dữ liệu đầu vào có thể gắn nhãn hoặc không gắn nhãn. Dựa trên dữ liệu đầu vào này thuật toán sẽ ước tính về mẫu dữ liệu từ đó đưa ra dự đoán hoặc phân loại.
* Hàm lỗi (Loss function): Loss function đánh giá dự đoán của mô hình. Nếu có các mẫu dữ liệu đã biết, loss function có thể so sánh để đánh giá độ chính xác của mô hình.
* Quy trình tối ưu hóa mô hình (Model optimization): Nếu mô hình có thể phù hợp hơn với các mẫu dữ liệu trong tập huấn luyện, thì các trọng số sẽ được điều chỉnh để giảm sự chênh lệch giữa dữ liệu đã biết và dữ liệu dự đoán của mô hình. Thuật toán sẽ lặp lại quy trình đánh giá, tối ưu hóa và cập nhật các trọng số của mô hình một cách tự động cho đến khi đạt đến ngưỡng tối ưu kỳ vọng.

Các thuật học máy được chia thành ba loại chính:

* Học máy có giám sát (Supervised machine learning): Supervised machine learning được xác định bằng cách sử dụng các bộ dữ liệu được gắn nhãn để huấn luyện các thuật toán nhằm phân loại dữ liệu hoặc dự đoán kết quả một cách chính xác.
* Học máy không giám sát (Unsupervised machine learning): Unsupervised machine learning sử dụng các thuật toán học máy để phân tích và phân cụm các bộ dữ liệu không được gắn nhãn. Các thuật toán này khám phá các mẫu hoặc nhóm dữ liệu ẩn mà không cần sự can thiệp của con người.
* Học máy bán giám sát (Semi-supervised machine learning): Trong quá trình đào tạo, semi-supervised machine learning sử dụng tập dữ liệu được gắn nhãn nhỏ hơn để huấn luyện khả năng phân loại phân loại của mô hình và trích xuất đặc trưng từ tập dữ liệu lớn hơn, không được gắn nhãn. Semi-supervised machine learning có thể giải quyết vấn đề không có đủ dữ liệu được gán nhãn cho thuật toán học có giám sát.

Hiện nay, đã có nhiều thuật toán học máy được nghiên cứu và phát triển ứng dụng.

Một số thuật toán học máy thường được sử dụng như: Linear regression, Logistic regression, Clustering, Decision trees, Random forests, Neural networks.

Một trong những điểm mạnh chính của học máy là khả năng xác định các mẫu và mối quan hệ phức tạp trong các tập dữ liệu lớn mà con người khó hoặc không thể phát hiện được [29]. Học máy cũng có khả năng tự động hóa nhiều nhiệm vụ hiện đang được thực hiện bởi con người, giải phóng thời gian và nguồn lực cho các nhiệm vụ phức tạp hoặc sáng tạo hơn.

Tuy nhiên, cũng có những thách thức liên quan đến học máy, đặc biệt là liên quan đến các vấn đề về sai lệch, công bằng và minh bạch. Các thuật toán học máy chỉ tốt khi dữ liệu được đào tạo đầy đủ và đúng đắn. Nếu dữ liệu này chứa các sai lệch hoặc không chính xác, điều này có thể dẫn đến các dự đoán sai lệch hoặc không chính xác. Vì vậy, những nghiên cứu và ứng dụng cần nhận thực được những thách thức này và hướng tới phát triển các thuật toán mạnh mẽ, công bằng và minh bạch.

## 2.4. Bài toán ứng dụng các thuật toán học máy vào mô hình chấm điểm tín dụng

### 2.4.1. Tổng quan

Áp dụng các thuật toán học máy vào mô hình chấm điểm tín dụng đề cập đến việc sử dụng các thuật toán học máy để phát triển các mô hình chấm điểm tín dụng chính xác và hiệu quả hơn [30]. Các mô hình chấm điểm tín truyền thống dựa trên các tính năng cố định và hạn chế, chẳng hạn như lịch sử tín dụng và hành vi thanh toán của người đi vay để tạo ra điểm tín dụng. Tuy nhiên, các thuật toán học máy có thể phân tích các tập dữ liệu lớn và phức tạp, bao gồm các nguồn dữ liệu phi truyền thống như hoạt động trên mạng xã hội, hành vi mua sắm trực tuyến, … để tạo ra điểm tín dụng chính xác hơn. Điều này có thể giúp các tổ chức tài chính đưa ra quyết định cho vay tốt hơn, cải thiện và quản lý rủi ro, đồng thời thúc đẩy sự ổn định và tăng trưởng tài chính.

Việc sử dụng các thuật toán học máy trong chấm điểm tín dụng đã thu hút được sự chú ý đáng kể trong những năm gần đây do khả năng cải thiện tính chính xác, công bằng và minh bạch của các mô hình chấm điểm tín dụng. Tuy nhiên, cũng có một số thách thức liên quan đến việc áp dụng các thuật toán học máy cho các mô hình chấm điểm tín dụng. Một trong những thách thức chính là vấn đề về khả năng diễn giải. Các thuật toán học máy như Neural Networks, Random Forest, … thường được coi là “black box”, nghĩa là hoạt động bên trong của thuật toán không thể nhìn thấy và khó hiểu. Điều này có thể dẫn đến những khó khăn trong việc giải thích quá trình ra quyết định của mô hình và có thể dẫn đến sự không tin tưởng từ cả người đi vay và người cho vay. Một số nghiên cứu đã đề xuất cách tiếp cận để giải quyết vấn đề này bằng cách kết hợp các thuật toán học máy khác nhau để tạo ra các mô hình dễ hiểu hơn.

Một thách thức khác là vấn đề thiên vị. Các thuật toán học máy chỉ khách quan khi dữ liệu mà chúng được đào tạo đầy đủ và chính xác. Nếu dữ liệu đào tạo bị sai lệch, mô hình kết quả cũng sẽ bị sai lệch. Điều này có thể dẫn đến việc đối xử không công bằng đối với một số nhóm người vay nhất định, chẳng hạn như những người đến từ các cộng đồng ít được đại diện. Những thách thức này cần phải được xem xét cẩn thận khi áp dụng các thuật toán học máy cho các mô hình chấm điểm tín dụng và phải thực hiện các biện pháp thích hợp để đảm bảo rằng các mô hình thu được là chính xác, công bằng và minh bạch.

Ngoài các vấn đề về sai lệch và minh bạch, vấn đề overfitting cũng là một thách thức khác liên quan đến các thuật toán học máy trong chấm điểm tín dụng. Overfitting xảy ra khi mô hình quá phức tạp và quá khớp với dữ liệu huấn luyện, dẫn đến khả năng dự đoán kém đối với dữ liệu mới. Overfitting có thể dẫn đến dự đoán không chính xác và giảm hiệu suất của mô hình. Để giải quyết vấn đề này, một số kỹ thuật sẽ được đề xuất trong báo cáo bao gồm chuẩn hóa dữ liệu và xác thực chéo (Cross-validation).

### 2.4.2. Nghiên cứu liên quan

Nghiên cứu được thực hiện bởi Jiang và cộng sự [31] đã so sánh độ chính xác của một số mô hình chấm điểm tín dụng khác nhau, bao gồm Logistic Regression, Decision Trees và Neural Networks. Nghiên cứu cho thấy mạng nơron (Neural Networks) là mô hình chính xác nhất để dự đoán rủi ro tín dụng, với tỷ lệ chính xác là 90%. Các tác giả gợi ý rằng Neural Networks có thể là một thuật toán hữu ích cho những tổ chức cho vay muốn cải thiện mô hình chấm điểm tín dụng của họ.

Trong một nghiên cứu của Bouzouita và các cộng sự [32] đã sử dụng phương pháp học máy để phát triển mô hình chấm điểm tín dụng cho các doanh nghiệp vừa và nhỏ (SME). Các tác giả nhận thấy rằng mô hình này có độ chính xác cao trong việc dự đoán rủi ro tín dụng cho các doanh nghiệp vừa và nhỏ, với tỷ lệ chính xác là 92,5%. Các tác giả gợi ý rằng mô hình này có thể là một công cụ hữu ích cho những người cho vay muốn đưa ra quyết định cho vay đúng đắn đối với các doanh nghiệp vừa và nhỏ.

Một thuật toán học máy phổ biến được sử dụng trong chấm điểm tín dụng là Decision Trees. Một nghiên cứu của Zhang và các cộng sự [33] đã đề xuất một mô hình sử dụng thuật toán Decision Trees để chấm điểm tín dụng từ một bộ dữ liệu lớn chứa rất nhiều thuộc tính, bao gồm cả dữ liệu tài chính truyền thống và dữ liệu phi truyền thống như sử dụng điện thoại di động và các hoạt động trên mạng xã hội. Nghiên cứu cho thấy mô hình sử dụng thuật toán Decision Trees hoạt động tốt hơn các mô hình tính điểm tín dụng truyền thống, với độ chính xác và điểm AUC cao hơn.

Ngoài ra, một nghiên cứu của Lachos-Perez và các cộng sự [34] đã đề xuất một mô hình chấm điểm tín dụng dựa trên lựa chọn và phân loại biến Bayes bằng cách sử dụng Logistic Regression và phân tích biệt thức tuyến tính. Mô hình đạt được độ chính xác cao và cho thấy những cải tiến đáng kể về khả năng diễn giải, tính công bằng và minh bạch. Các tác giả cũng nhấn mạnh về tầm quan trọng của khả năng diễn giải và tính minh bạch trong các mô hình chấm điểm tín dụng, đặc biệt là trong việc tránh thiên vị và phân biệt đối xử đối với một số nhóm người đi vay.

Việc sử dụng các thuật toán học máy trong chấm điểm tín dụng đã cho thấy kết quả đầy hứa hẹn trong việc cải thiện tính chính xác, công bằng và minh bạch của các mô hình chấm điểm tín dụng. Random Forest, Gradient Boosted Decision Trees, Neural

Networks là các thuật toán học máy phổ biến được sử dụng để chấm điểm tín dụng. Các thuật toán này sử dụng các nguồn dữ liệu truyền thống và phi truyền thống để tạo điểm tín dụng có thể cung cấp cái nhìn sâu sắc hơn về mức độ tin cậy của người đi vay.

## 2.5. Thuật toán học máy được ứng dụng vào xây mô hình chấm điểm tín dụng

### 2.5.1. Logistic regression

#### 2.5.1.1. Tổng quan

Logistic regression là một thuật toán học máy có giám sát được sử dụng cho các tác vụ phân loại, logistic regression dự đoán xác suất của một biến mục tiêu [35]. Thuật toán logistic regression mô hình hóa mối quan hệ giữa các thuộc tính dữ liệu đầu vào và phân lớp đầu ra bằng cách sử dụng hàm logistic, còn được gọi là hàm sigmoid.

Hàm logistic nhận bất kỳ số có giá trị thực nào và đầu ra là giá trị trong khoảng (0,1). Thuật toán học các tham số của hàm logistic bằng cách giảm thiểu hàm chi phí, điển hình là cross-entropy loss với việc sử dụng thuật toán tối ưu hóa như gradient descent. Thuật toán logistic regression có thể xử lý cả tính năng đầu vào phân loại và liên tục và thường được sử dụng trong các lĩnh vực khác nhau, bao gồm tài chính, chăm sóc sức khỏe và tiếp thị, …. Thuật toán này dễ diễn giải và có thể cung cấp thông tin chi tiết về mối quan hệ giữa các tính năng đầu vào và lớp đầu ra.

#### 2.5.1.2. Lý thuyết liên quan

Đầu ra dự đoán của logistic regression:

𝑓(𝐱) = 𝜃(𝐰𝑇𝐱) *PT 2.1*

Trong đó:

* 𝐱 là dữ liệu đầu vào.
* 𝜃 là hàm logistic.
* 𝐰là các tham số của thuật toán.

Đầu ra dự đoán của logistic regression là giá trị xác suất của biến mục tiêu thuộc về lớp positive (lớp dữ liệu quan trọng hơn cần được xác định đúng của bài toán), với các thuộc tính dữ liệu đầu vào. Hàm logistic, còn được gọi là hàm sigmoid, được sử dụng để chuyển đổi tổ hợp tuyến tính của các thuộc tính đầu vào thành giá trị xác suất trong khoảng (0,1). Công thức hàm sigmoid được định nghĩa như sau:

1 *PT 2.2*

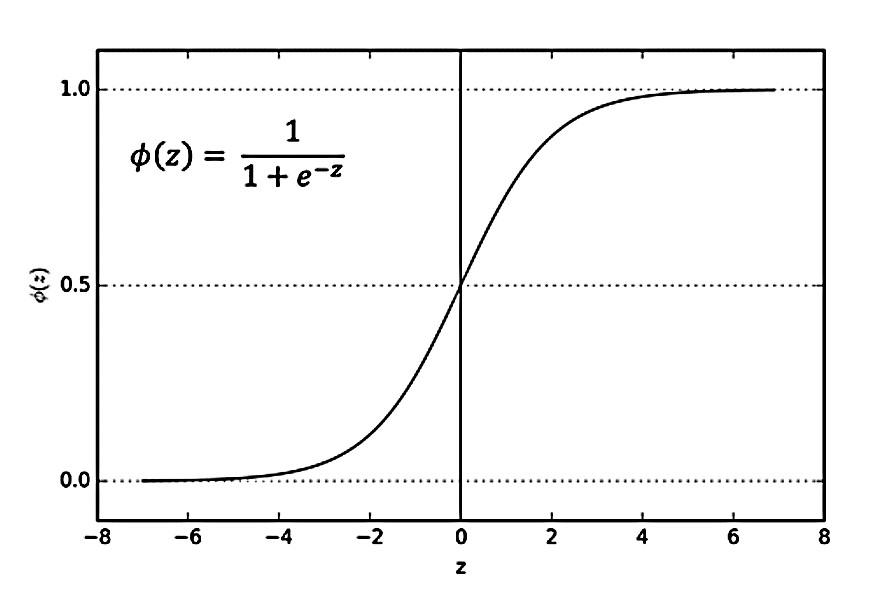
𝜙(𝑧) = 1 + 𝑒−𝑧

Ngoài ra, hàm tanh cũng có thể được sử dụng được sử dụng:

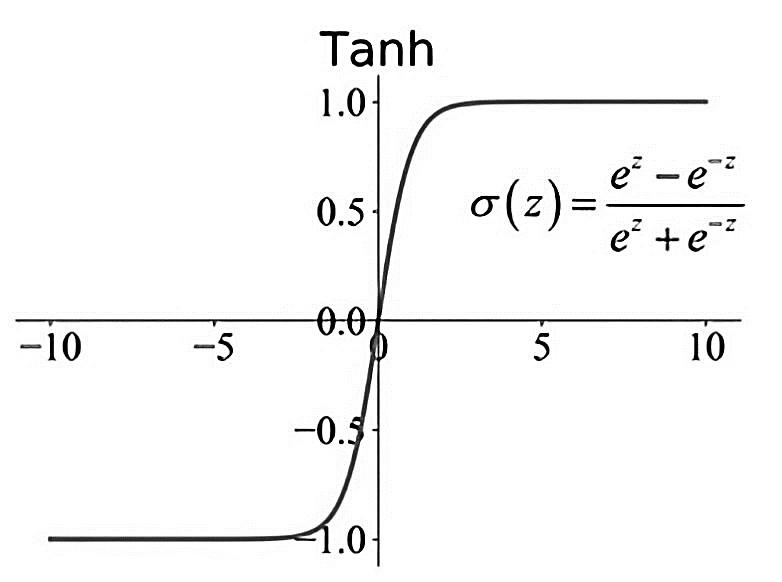
𝑒𝑧 − 𝑒−𝑧 *PT 2.3*

tanh(𝑧) = 𝑓(𝑧) = 𝑒𝑧 + 𝑒−𝑧

𝑓′(𝑧) = 1 − tanh2(𝑧) *PT 2.4*



*Hình 2.1 Biểu diễn đồ thị hàm Sigmoid*



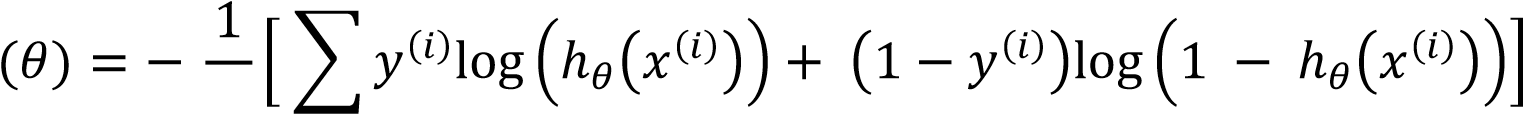
*Hình 2.2 Biểu diễn đồ thị hàm Tanh*

Hàm Tanh nhận giá trị trong khoảng (−1, 1) nhưng các giá trị có thể dễ dàng đưa về khoảng (0, 1) để phù hợp với thuật toán.

Hàm mất mát của thuật toán logistic regression được sử dụng để đo lỗi hoặc sự khác biệt giữa đầu ra dự đoán và đầu ra thực tế đối với tập hợp các tham số của mô hình. Trong logistic regression, mục tiêu là tối đa hóa khả năng xảy ra của dữ liệu huấn luyện với tham số mô hình. Hàm mất mát được sử dụng cho hồi quy logistic là log-likelihood, còn được gọi là hàm mất mát cross-entropy [36].

Công thức cho hàm mất mát của thuật toán hồi quy logistic:

𝒥 *PT 2.5*

𝑚

Trong đó:

* 𝒥(𝜃): Hàm mất mát.
* 𝜃: Các tham số của mô hình.
* 𝑚: Số lượng mẫu dữ liệu.
* 𝑦(𝑖): Giá trị đầu ra thực tế của mẫu dữ liệu 𝑥(𝑖).
* ℎ𝜃(𝑥(𝑖)): Giá trị đầu ra dự đoán của mẫu dữ liệu 𝑥(𝑖).

Hàm mất mát tính toán sự khác biệt giữa giá trị đầu ra được dự đoán và giá trị đầu ra thực tế cho từng mẫu đào tạo, tổng hợp trên tất cả các mẫu đào tạo. Mục tiêu của thuật toán logistic regression là tìm các giá trị của 𝜃 nhằm tối ưu hàm mất mát. Quá trình này thường sử dụng thuật toán tối ưu giảm độ dốc ngẫu nhiên (SGD). Công thức tổng quát của thuật toán SGD được đưa ra như sau:

𝜃𝑗 = 𝜃𝑗 − 𝛼 (𝑦(𝑖) − ℎ𝜃(𝑥(𝑖) )) 𝑥𝑗(𝑖) *PT 2.6*

Trong đó:

* 𝜃𝑗: Tham số thứ 𝑗 được cập nhật.
* 𝛼: Tốc độ học (LR).
* 𝑦(𝑖): Giá trị đầu ra thực tế của mẫu dữ liệu huấn luyện thứ 𝑖.
* ℎ𝜃(𝑥(𝑖)): Giá trị đầu ra dự đoán của mẫu dữ liệu huấn luyện thứ 𝑖.
* 𝑥𝑗(𝑖): Thuộc tính thứ 𝑗 của mẫu dữ liệu đào tạo thứ 𝑖.

Công thức trên tính toán sự khác biệt giữa các giá trị đầu ra được dự đoán và giá trị đầu ra thực tế cho một mẫu dữ liệu đào tạo duy nhất và cập nhật từng tham số theo giá trị thuộc tính tương ứng của nó và sự khác biệt giữa các giá trị đầu ra được dự đoán và giá trị đầu ra thực tế. Tốc độ học 𝛼 kiểm soát kích thước bước giảm của quá trình cập nhật tham số, với giá trị 𝛼 lớn quá trình hội tụ của mô hình sẽ diễn ra nhanh hơn, nhưng cũng có nguy cơ vượt quá giá trị tham số tối ưu. Thuật toán SGD áp dụng quá trình lặp lại công thức cập nhật này cho từng mẫu dữ liệu huấn luyện trong tập dữ liệu, cho đến khi đạt được sự hội tụ hoặc số lần lặp tối đa.

Bằng cách tối ưu hàm mất mát, thuật toán có thể tìm được tập hợp tham số mô hình tốt nhất để tối đa hóa khả năng dự đoán của mô hình thông qua dữ liệu huấn luyện từ đó cải thiện độ chính xác của các dự đoán trên dữ liệu mới.

### 2.5.2. Random forest

#### 2.5.2.1. Tổng quan

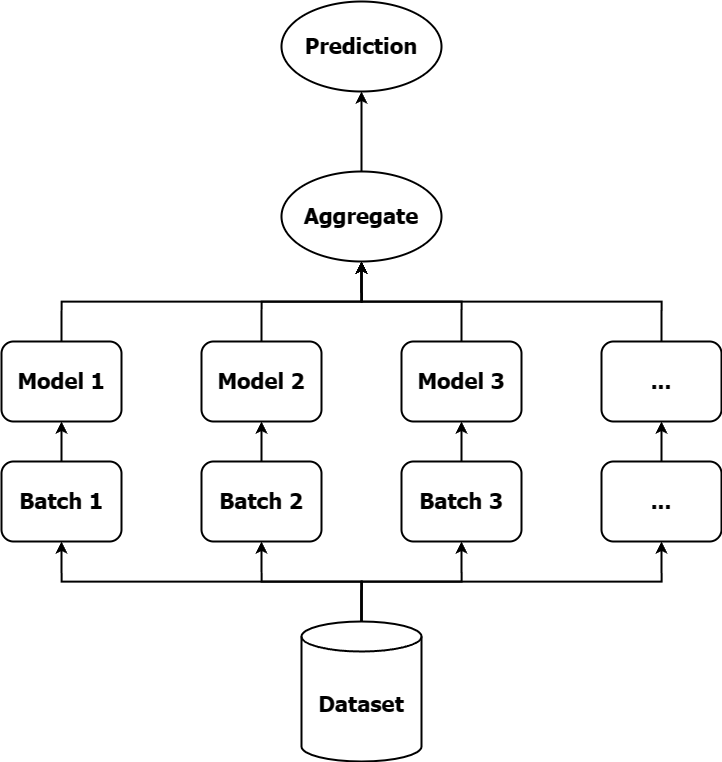
Random forest là một thuật học có giám sát được sử dụng phổ biến trong học máy. Thuật toán này dựa trên thuật toán cây quyết định (decision trees), là thuật toán phân loại dữ liệu dựa trên một loạt các quyết định phân loại. Thuật random forest xây dựng nhiều cây quyết định và kết hợp chúng để cải thiện độ chính xác của mô hình.

Ý tưởng cơ bản của thuật toán random forest là tạo ra một số lượng lớn cây quyết định bằng cách chọn ngẫu nhiên một tập hợp con các thuộc tính và huấn luyện từng cây trên một tập hợp con ngẫu nhiên của dữ liệu huấn luyện. Mỗi cây trong cụm sẽ dự đoán nhãn lớp của một đầu vào nhất định và dự đoán cuối cùng được thực hiện bằng cách kết hợp các dự đoán của tất cả các cây đã được xây dựng. Quá trình này được gọi là học đồng bộ (ensemble learning), trong đó nhiều mô hình được sử dụng để đưa ra dự đoán chính xác hơn các mô hình riêng lẻ [37] [38].

Thuật toán random forest có một số lợi thế so với các thuật toán học máy khác. Thuật toán có thể xử lý dữ liệu nhiều chiều, dữ liệu nhiễu và các giá trị bị thiếu và cũng có thể cung cấp các ước lượng về tầm quan trọng của thuộc tính. Hơn nữa, nó ít bị overfitting hơn so với một cây quyết định duy nhất vì có xu hướng ghi nhớ dữ liệu huấn luyện [39].

#### 2.5.2.2. Lý thuyết liên quan

Học đồng bộ (ensemble learning) là một kỹ thuật trong đó kết hợp nhiều mô hình riêng lẻ để tạo ra một mô hình dự đoán mạnh mẽ và chính xác hơn. Ensemble learning được sử dụng rộng rãi trong học máy vì chúng có thể cải thiện độ chính xác và hiệu suất của các mô hình riêng lẻ. Ensemble learning có thể giúp giảm việc trang bị quá mức (overfitting) và nắm bắt các mẫu đa dạng trong dữ liệu từ đó cho ra mô hình dự đoán mạnh mẽ và chính xác hơn [40] [41] [42].

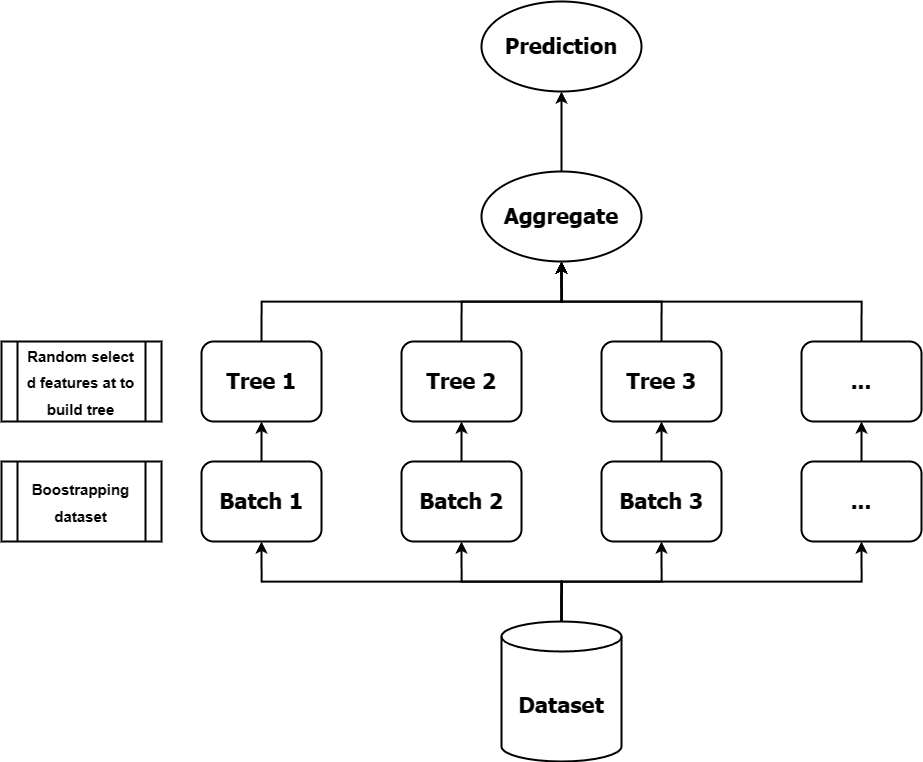


*Hình 2.3 Kiến trúc tổng quan về mô hình kết hợp*

Lấy mẫu tái lập (bootstrap) phương pháp lấy mẫu thống kê được sử dụng trong học máy, có liên quan chặt chẽ với thuật toán random forest. Kỹ thuật này liên quan đến việc lấy mẫu dữ liệu với sự thay thế từ tập dữ liệu gốc để tạo tập dữ liệu mới, sau đó được sử dụng để huấn luyện nhiều mô hình. Phương pháp lấy mẫu tái lập tạo ra nhiều bộ dữ liệu tương tự như dữ liệu ban đầu, nhưng có chút thay đổi nhỏ, nhằm giảm tình trạng overfitting và cải thiện độ chính xác của mô hình [43] [44] [45] [46].

Mô hình random forest sẽ áp dụng cả hai phương pháp ensemble learning và bootstrap. Mỗi cây được đào tạo trên một mẫu bootstrap khác nhau của dữ liệu gốc.

Mẫu bootstrap được lấy bằng phương pháp bootstrap đã đề cập ở phần trên kết quả của quá trình này là một tập dữ liệu mới có cùng số lượng quan sát như tập dữ liệu gốc, nhưng với một số quan sát được lặp lại và những quan sát khác bị loại trừ.



*Hình 2.4 Quá trình tạo thành một mô hình của thuật toán Random forest*

Các bước tạo mô hình từ thuật toán Random forest:

1. Lấy 𝐍 mẫu ngẫu nhiên với từ tập dữ liệu gốc để tạo mẫu bootstrap.
2. Huấn luyện mô hình decision trees trên mẫu bootstrap bằng cách sử dụng ngẫu nhiên 𝐝 thuộc tính.
3. Lặp lại bước 1 và bước 2 để xây dựng 𝐌 mô hình decision trees.
4. Khi đã xây dựng đủ số lượng mô hình decision trees đã đặt ra. Tiến hành dự đoán giá trị biến mục tiêu bằng cách tổng hợp các dự đoán của tất cả 𝐌 mô hình decision trees với phương pháp lấy giá trị trung bình của các dự đoán (đối với mô hình dự báo) hoặc bằng cách sử dụng biểu quyết đa số (đối với mô hình phân loại).

Kết quả dự đoán từ mô hình random forest là sự kết hợp của nhiều mô hình decision trees nên sẽ giúp cải thiện độ chính xác so với chỉ sử dụng một mô hình decision trees. Đồng thời giúp cho kết quả ít bị chệch, giảm thiểu được hiện tượng overfitting ở mô hình decision trees, một điều mà mô hình decision trees thường xuyên gặp phải [47].

### 2.5.3. XGBoost

#### 2.5.3.1. Tổng quan

XGBoost (Extreme Gradient Boosting) [48] là một thuật toán học máy phổ biến được sử dụng rộng rãi cho các bài toán hồi quy và phân loại. Thuật toán sử dụng tăng cường độ dốc (gradient boosting) và được thiết kế để có hiệu quả cao và có thể mở rộng, làm cho nó trở nên lý tưởng cho các bộ dữ liệu lớn. Gradient boosting là một thuật toán học có giám sát, cố gắng dự đoán chính xác một biến mục tiêu bằng cách kết hợp các ước tính của một tập hợp các mô hình yếu hơn, đơn giản hơn [49].

XGBoost hoạt động bằng cách tạo một chuỗi các cây quyết định (decision trees), trong đó mỗi cây tiếp theo được tạo để sửa lỗi của các cây trước đó. Thuật toán có nhiều siêu tham (hyperparameters) số có thể được điều chỉnh để tối ưu hóa hiệu suất của thuật toán, bao gồm tốc độ học, số lượng cây và độ sâu tối đa của mỗi cây.

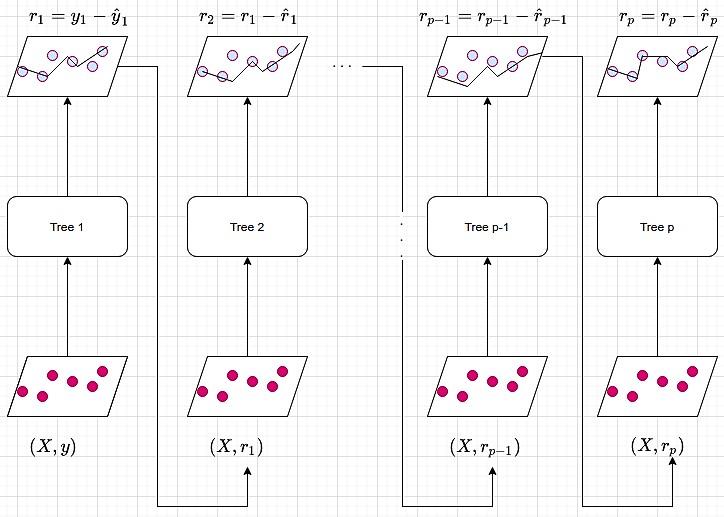
XGBoost đã trở thành một thuật toán phổ biến trong nhiều lĩnh vực khác nhau, bao gồm tài chính, chăm sóc sức khỏe, tiếp thị, …. Nó đã được sử dụng để giải quyết các vấn đề phức tạp, chẳng hạn như dự đoán các giao dịch gian lận, chẩn đoán tình trạng y tế và xác định khả năng rời bỏ của khách hàng. Tính linh hoạt, độ chính xác và khả năng mở rộng đã khiến XGBoost trở thành một trong những thuật toán học máy được sử dụng rộng rãi nhất hiện nay.

#### 2.5.3.2. Lý thuyết liên quan

XGBoost là thuật toán với mục đích đưa ra một mô hình dự đoán dưới dạng một tập hợp các mô hình dự đoán yếu, thường là các cây quyết định (decision trees) [50] [51].

Thuật toán này được các nghiên cứu đánh giá có hoạt động vượt trội hơn so với thuật toán random forest [50] [51] [52].

Thuật toán không sử dụng sai số của mô hình để tính toán trọng số cho dữ liệu mà sử dụng phần dư. Với mô hình tập hợp từ các mô hình cây quyết định, Mỗi cây quyết định sẽ được thành lập phụ thuộc vào kết quả dự báo của cây quyết định liền trước. Tại một cây quyết định mô hình sẽ tìm cách khớp phần dư từ cây quyết định trước đó.



*Hình 2.5 Huấn luyện mô hình XGBoost*

Với dữ liệu đầu vào 𝐗 và biến mục tiêu là 𝒚, thuật toán gradient boosting cố gắng tạo ra hàm 𝒇̂(𝒙) với mục tiêu dự đoán. Tại mô hình thứ 𝒃 trong chuỗi các mô hình dự đoán hàm mục tiêu tại mô hình đó là 𝒇̂𝒃. Mô hình tìm cách khớp giá trị phần dư 𝐫𝐢 từ cây quyết định trước là 𝒇̂𝒃−𝟏. Quá trình được mô tả như sau:

1. Thiết lập hàm mục tiêu 𝒇̂(𝒙) = 𝟎và phần dư 𝐫𝟎 = 𝒚.
2. Lặp lại quá trình huấn luyện cây quyết định theo chuỗi tương ứng với 𝑏 = 1,2,3, … , 𝐵. Với các bước nhỏ:
   1. Khớp cây quyết định 𝒇̂𝒃 có độ sâu cây là 𝒅 trên tập huấn luyện (𝐗, 𝐫𝒃).
   2. Cập nhật 𝒇 bằng cách cộng thêm vào giá trị dự báo của một cây quyết định, giá trị này được nhân với hệ số co 𝝀(Hệ số này gần giống như learning rate có tác dụng kiểm soát tỷ lệ mà gradient boosting cập nhật

|  |  |  |
| --- | --- | --- |
| phần dư): |  |  |
|  | 𝒇̂(𝐱) = 𝒇̂(𝐱) + 𝝀𝒇̂𝑏(𝐱) | *PT 2.13* |

|  |  |
| --- | --- |
| 3. Cập nhật phần dư cho mô hình: |  |
| 𝐫 ≔ 𝐫 − 𝝀𝒇̂𝒃(𝐱) | *PT 2.14* |

𝑏+1 𝑏

1. Kết quả dự báo sẽ là kết quả kết hợp từ các mô hình con:

𝐵 *PT 2.15*

𝑓̂(𝐱) = ∑   𝜆𝑓̂𝑏(𝐱)

𝑏=1

XGBoost cải thiện thuật toán tăng cường độ dốc (gradient boosting) về mặt tốc độ và quy mô tính toán [53]. XGBoost sử dụng nhiều nhân CPU để quá trình học có thể diễn ra song song trong lúc đào tạo. XGBoost là một thuật toán tăng cường có thể xử lý các tập dữ liệu mở rộng, khiến nó trở nên hấp dẫn đối với những ứng dụng dữ liệu lớn.

## 2.6. Kỹ thuật và phương pháp sử dụng trong nghiên cứu và thực nghiệm

### 2.6.1. Feature engineering

Feature engineering là một kỹ thuật học máy tận dụng dữ liệu để tạo các biến mới không có trong tập huấn luyện. Nó có thể tạo ra các tính năng mới cho cả mô hình học có giám sát và không giám sát, với mục tiêu đơn giản hóa và tăng tốc độ chuyển đổi dữ liệu, đồng thời nâng cao độ chính xác của mô hình [54].

Các kỹ thuật trong feature engineering được sử dụng trong quá trình nghiên cứu và thực nghiệm bao gồm: Imputation, Feature extraction, One-hot encoding, Feature scaling, Combine.

#### 2.6.1.1. Imputation

Imputation là quá trình điền vào các giá trị còn thiếu trong tập dữ liệu. Có nhiều cách để điền dữ liệu bị thiếu, có thể điền giá trị trung bình, trung vị hoặc sử dụng các kỹ thuật imputation nâng cao như K-nearest neighbor (KNN) hoặc hồi quy (regressionbased). Kỹ thuật imputation ngăn ngừa việc mất mát dữ dữ liệu, giúp cho mô hình huấn luyện tiếp nhận được đầy đủ thông tin trong quá trình huấn luyện từ đó cải thiện chất lượng huấn luyện của mô hình.

#### 2.6.1.2. Feature extraction

Feature extraction sử dụng các thuộc tính hiện có trong dữ liệu thô để tạo các thuộc tính mới cho quá trình huấn luyện mô hình. Điều này liên quan đến việc chọn các thông tin có liên quan từ dữ liệu thô và chuyển đổi thành một số biểu diễn hoặc thông tin mới có ý nghĩa hơn để mô hình nắm bắt được tổng quát hơn về dữ liệu. Ví dụ như trích xuất “tỷ lệ khoản vay” từ hai thuộc tính là “khoản vay tín dụng yêu cầu” và “khoản vay tín dụng được cấp”.

Để định lượng khả năng dự đoán của một thuộc tính trong quá trình trích xuất đặc trưng thì trọng số dấu hiệu (weight of evidence – WOE) và chỉ số giá trị thông tin (information value - IV) được sử dụng nhằm tính toán lượng thông tin mà một thuộc tính cung cấp về biến mục tiêu trong tập dữ liệu [55] [56] [57].

WOE (weight of evidence) là một trong những kĩ thuật feature engineering được áp dụng trong việc đánh giá chất lượng của thuộc tính dữ liệu. Trong mô hình dự đoán rủi ro tín dụng, WOE được tính toán dựa trên tỷ lệ khách hàng xấu (khách hàng không trả hoặc trả chậm khoản vay) và khách hàng tốt (khách hàng hoàn trả được khoản vay):

% of non − events *PT 2.16*

𝐖𝐎𝐄 = 𝐥𝐧 ( )

% of events

Với "% of non − events" là phân phối phi sự kiện trong dữ liệu (với dữ liệu cho mô hình rủi ro tín dụng sẽ là phân phối khách hàng tốt), "% of events" là phân phối của các sự kiện trong dữ liệu (khách hàng xấu).

Các bước tính chỉ số WOE:

* Bước 1: Đối với biến liên tục sẽ chia dữ liệu thành k nhóm (bin). Đối với các biến phân loại sẽ bỏ qua bước này.
* Bước 2: Tính số lượng "events" và "non − events" trong mỗi bin.
* Bước 3: Tính "% of non − events" và "% of events" trên trong mỗi nhóm.
* Bước 4: Tính chỉ số WOE bằng công thức ở phương trình *PT 2.16*

Các quy tắc liên quan:

* Số lượng mẫu trong bin phải lớn hơn 5% trên tổng lượng mẫu dữ liệu.
* Số lượng "events" và "non − events" phải khác nhau trong trong mỗi bin.

Từ chỉ số WOE ta có thể tính được chỉ số IV. IV là chỉ số rất hữu ích trong việc chọn lọc các biến quan trọng trong mô hình dự đoán. IV được tính theo công thức:

% of non − events *PT 2.17*

𝐈𝐕 = ∑( ) × 𝐖𝐎𝐄

% of events

*Bảng 2.1 Bảng phân chia giá trị thông tin của các thuộc tính dựa trên khoảng giá trị của chỉ số IV*

|  |  |
| --- | --- |
| **Giá trị thông tin (IV)** | **Khả năng dự đoán của thuộc tính** |
| Nhỏ hơn 0.02 | Thuộc tính không hữu ích cho mô hình |
| 0.02 – 0.1 | Khả năng dự đoán yếu |
| 0.1 – 0.3 | Khả năng dự đoán trung bình |
| 0.3 – 0.5 | Khả năng dự đoán mạnh |
| Lớn hơn 0.5 | Thuộc tính đáng ngờ (Cần kiểm tra) |

Chỉ số trọng số dấu hiệu (WOE) và chỉ số giá trị thông tin (IV) là thước đo hữu ích trong việc lựa chọn và trích xuất đặc trưng. Những chỉ số này cung cấp thước đo định lượng về khả năng dự đoán thuộc tính dữ liệu. Bằng cách chọn các tính năng có giá trị thông tin cao nhất, có thể xây dựng các mô hình dự đoán chính xác và hiệu quả hơn.

### 2.6.2. Gradient descent

Gradient descent là một thuật toán tối ưu được sử dụng để giảm thiểu hàm mất mát (loss function) trong các mô hình học máy. Mục tiêu của thuật toán này là tìm giá trị của các tham số mô hình mà ở đó sẽ tối thiểu được loss function, bằng cách điều chỉnh lặp đi lặp lại các giá trị tham số theo hướng dốc nhất của đạo hàm loss function [58].

Giả sử, trong trường hợp mô hình logistic regression, loss function của mô hình logistic regression sẽ là:

𝑚

𝐽(𝜃) = 1 ∑ LCE(𝑓(𝑥(𝑖); 𝜃), 𝑦(𝑖)) 𝑚

𝑖=1

*PT 2.18*

− log (𝑓(𝑥(𝑖); 𝜃)) nếu 𝑦(𝑖) = 1

LCE(ℎ𝜃(𝑥(𝑖)), 𝑦(𝑖)) = { − log (1 − 𝑓(𝑥(𝑖); 𝜃)) nếu 𝑦(𝑖) = 0

Mục tiêu của gradient descent là tìm được bộ tham số 𝜃 để tối ưu loss function:

𝜃̂ = argmin 1 𝑚 (𝑖); 𝜃), 𝑦(𝑖))  *PT 2.19*

∑ LCE(𝑓(𝑥

𝜃 𝑚 𝑖=1

Thuật toán Gradient descent bắt đầu với việc lấy ngẫu nhiên cho giá trị tham số 𝜃 và lặp lại cập nhật các giá trị tham số theo hướng đạo hàm mang giá trị âm của loss function, quá trình giảm này sử dụng thêm chỉ số 𝛼 được gọi là tốc độ học (learning rate) nhằm tăng tốc độ giảm. Công thức để cập nhật các tham số:

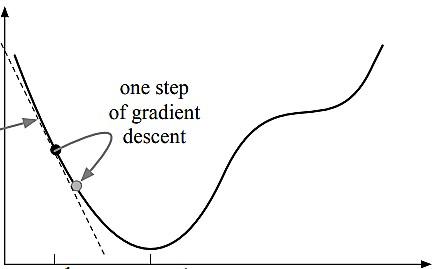
𝜕𝐽(𝜃) *PT 2.20*

𝜃𝑖 = 𝜃𝑖 − 𝛼

𝜕𝜃𝑖

Trong đó:

* 𝜃𝑖 đại diện cho tham số thứ 𝑗 được cập nhật.
* 𝜕𝐽𝜕(𝜃𝜃𝑖) là đạo hàm riêng (gradient - ∇) của loss function đối với tham số 𝜃𝑗.



*Hình 2.6 Minh họa một bước giảm của loss function [59]*

Các bước tối ưu tham số 𝜃 của thuật toán gradient descent [59]:

1. Lấy ngẫu nhiên tham số 𝜃0 khi đó 𝜃 = 𝜃0.
2. Quá trình lặp bao gồm:
   1. Xác định ∇𝜃𝑖 = ∇𝜃LCE(𝑓(𝑥(𝑖); 𝜃), 𝑦(𝑖)).
   2. Cập nhật 𝜃𝑖.
3. Dừng vòng lặp khi giá trị được cập nhật không còn thay đổi hoặc khoảng cách giữa 2 giá trị liên tiếp đủ nhỏ khi đó:

‖𝜃𝑖 𝜀 *PT 2.21*

Sau quá trình thực hiện thuật toán gradient descent sẽ thu được bộ tham số 𝜃 mà tại đó loss function đạt giá trị tối ưu nhất.

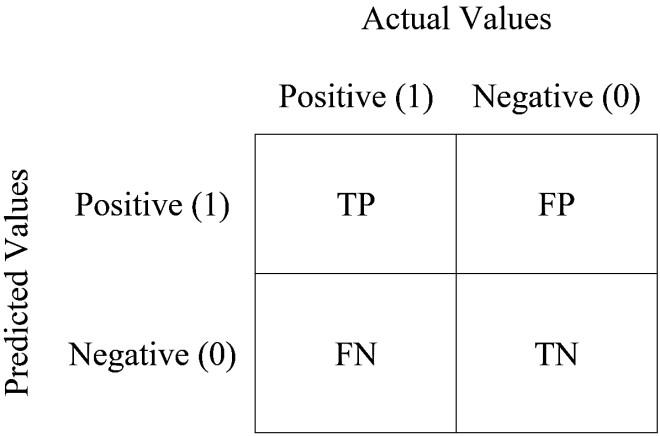
### 2.6.3. Phương pháp đánh giá mô hình

#### 2.6.3.1. Confusion matrix

Trong học máy, các mô hình phân loại được sử dụng để phân chia dữ liệu thành các danh mục khác nhau. Theo Jianfeng và các cộng sự [60], Confusion matrix là phương pháp để tóm tắt hiệu suất của mô hình phân loại. Nó là một công cụ cần thiết để đánh giá độ chính xác của một mô hình và xác định mức độ phân loại dữ liệu, đặc biệt là các mô hình đánh giá rủi ro tín dụng khi mà các dữ liệu thường có xu hướng mất cân bằng rất lớn [61] [62] khiến cho kết quả của chỉ số đánh giá độ chính xác phân loại (classification accuracy) không còn đánh giá đúng chất lượng của mô hình huấn luyện [63].

Giả sử, với mô hình phân loại khách hàng rủi ro, nếu xem khách hàng rủi ro là “positive” và khách hàng không có rủi ro được xem là “negative”. Quá trình xây dựng confusion matrix cho mô hình phân loại 2 lớp bao gồm các bước sau:

1. Tạo tập dữ liệu kiểm thử với các biến mục tiêu phân loại ứng với các mẫu dữ liệu.
2. Đưa ra dự đoán cho từng mẫu dữ liệu trong tập dữ liệu kiểm thử bằng đầu ra của mô hình phân loại đã được huấn luyện.
3. Từ kết quả thật và kết quả dự đoán của các dữ liệu trong dữ liệu kiểm thử.
4. Đưa ra số lượng cho các lớp TP, TN, FP, FN. Trong đó:
   * TP (True Positive): Số lượng mẫu dữ liệu được mô hình dự đoán là “positive” và thực tế nhãn của mẫu dữ liệu là “positive”.
   * TN (True Negative) : Số lượng mẫu dữ liệu được mô hình dự đoán là “negative” và thực tế nhãn của mẫu dữ liệu là “negative”.
   * FP (False Positive): Số lượng mẫu dữ liệu được mô hình dự đoán là “positive” tuy nhiên nhãn thực tế của mẫu là “negative”. (Những trường hợp dự đoán này được phân loại là sai lầm loại I).
   * FN (False Negative): Số lượng mẫu dữ liệu được mô hình dự đoán là “negative” tuy nhiên nhãn thực tế của mẫu là “positive” (Những trường hợp dự đoán này được phân loại là sai lầm loại II).



*Hình 2.7 Confusion Matrix*

Để biết mức độ chính xác của mô hình sẽ cần thêm một số chỉ số nhằm xác định hiệu suất phân loại của mô hình thông qua các chỉ số TP, TN, FP, FN. Các chỉ số đánh giá liên quan đến confusion matrix:

**Recall (Sensitivity - TPR)**: Tỷ lệ của các kết quả dự đoán là “positive” trên tổng số mẫu “positive” thực tế. Trong mô hình phân loại sẽ kỳ vọng chỉ số này lớn nhất.

TP PT 2.26

Recall = TPR =

TP + FN

**Precision**: Tỷ lệ của các kết quả dự đoán “positive” là đúng trên tổng số mẫu dự đoán là “positive”. Trong mô hình phân loại sẽ kỳ vọng chỉ số này lớn nhất.

TP *PT 2.27*

Precision = TP + FP

**Accuracy**: Tỷ lệ các dự đoán chính xác trên tổng số mẫu dự đoán.

TP + TN *PT 2.28*

Accuracy =

TP + TN + FP + FN

**F1 score**: Sử dụng “harmonic mean” để tính toán giữa 2 chỉ số recall và precision giúp việc so sánh giữa 2 mô hình có recall thấp, precision cao và ngược lại được thuận lợi hơn.

2 × Recall × Precision *PT 2.29*

F1 score =

Recall + Precision

**FPR**: Tỷ lệ dự đoán sai các trường hợp “positive” trên tổng số mẫu dự đoán là “negative”.

FP *PT 2.30*

FPR =

FP + TN

**Specificity**: Tỷ lệ dự đoán đúng các trường hợp “negative” trên tổng số trường hợp

“negative” thực tế

TN *PT 2.31*

Specificity = = 1 − FPR

FP + TN

Confusion matrix cung cấp trực quan kết quả dự đoán của mô hình. Một trong những lợi ích chính của việc sử dụng confusion matrix là nó giúp xác định điểm mạnh và điểm yếu của mô hình. Bằng cách xem xét confusion matrix, có thể biết mô hình đang mắc phải loại lỗi nào và điều chỉnh cách tiếp cận phù hợp. Ví dụ: nếu chỉ số FP quá cao, thì cần điều chỉnh ngưỡng dự đoán để giảm số lượng dự đoán FP.

#### 2.6.3.2. AUC-ROC curve

AUC-ROC curve là một phương pháp tính toán hiệu suất của một mô hình phân loại theo các ngưỡng phân loại khác nhau. Với bài toán phân loại nhị phân kết quả đầu ra của mô hình sẽ là xác suất trong khoảng (0,1) và việc chọn ngưỡng để phân lớp đầu ra này rất quan trọng.

Đường cong ROC (ROC curve) đồ thị của tỷ lệ TPR (*PT 2.26*) so với tỷ lệ FPR (*PT 2.30*) với các ngưỡng phân loại khác nhau. Để tạo ROC curve, xác suất dự đoán của mô hình cho lớp “positive” được sắp xếp theo thứ tự giảm dần. Sau đó, một ngưỡng dự đoán được đặt ra và tất cả các trường hợp có xác suất dự đoán lớn hơn hoặc bằng ngưỡng đó và được phân loại là “positive”, trong khi các trường hợp khác được phân loại là “negative”. TPR và FPR được tính toán dựa trên các phân loại này và quy trình được lặp lại cho các ngưỡng khác nhau [64] [65].

AUC là vùng bên dưới đường cong ROC, biểu thị xác suất mà một trường hợp

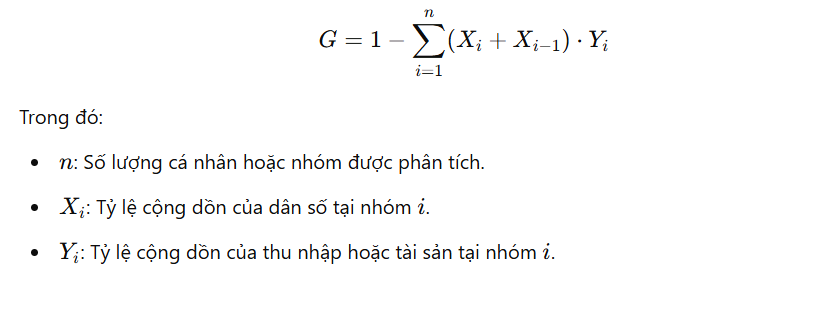
“positive” được chọn ngẫu nhiên sẽ được xếp hạng cao hơn một trường hợp “negative” được chọn ngẫu nhiên theo mô hình. AUC nằm trong khoảng [0,1], với AUC = 0,5 cho biết mô hình không tốt và hoàn toàn không có khả năng phân loại giữa 2 lớp, AUC = 1 cho biết mô hình phân loại rất tốt [66].

#### 2.6.3.3. Gini

Hệ số Gini (Gini Coefficient) là một thước đo kinh tế được sử dụng để đánh giá mức độ bất bình đẳng trong phân phối thu nhập hoặc tài sản trong một quần thể. Hệ số này do nhà kinh tế học người Ý Corrado Gini đề xuất vào năm 1912.

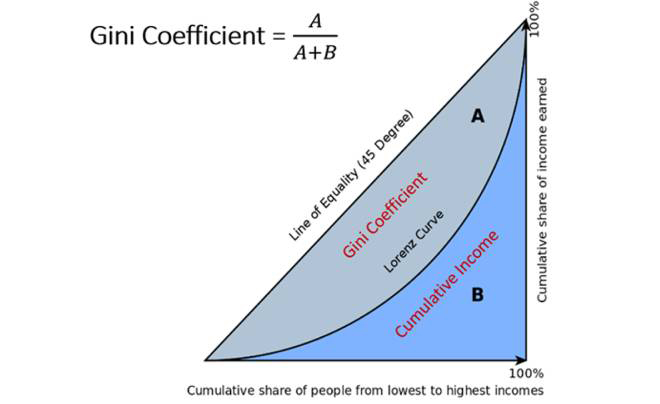
Hệ số Gini có giá trị nằm trong khoảng từ 0 đến 1:

* **0**: Thể hiện sự bình đẳng hoàn hảo (mọi người đều có thu nhập hoặc tài sản như nhau).
* **1**: Thể hiện sự bất bình đẳng tuyệt đối (một người sở hữu toàn bộ thu nhập hoặc tài sản, trong khi những người khác không có gì).



*Hình 2.8 Công thức tính hệ số Gini*

Hệ số Gini thường được biểu diễn qua **đường cong Lorenz**, thể hiện mối quan hệ giữa tỷ lệ cộng dồn của dân số (trục hoành) và tỷ lệ cộng dồn của thu nhập hoặc tài sản (trục tung).



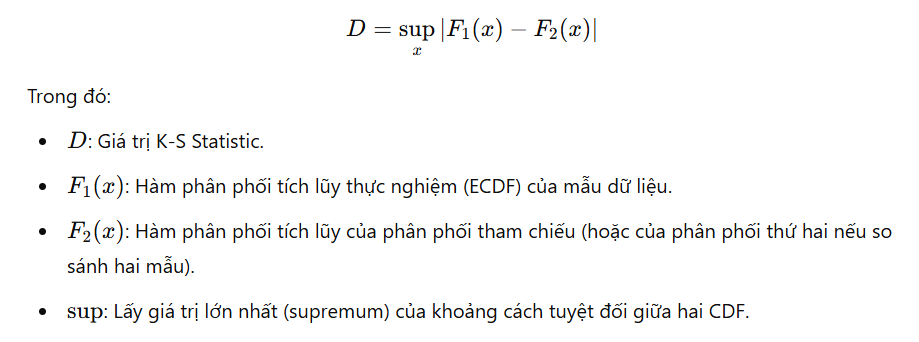
*Hình 2.9 Biểu đồ đường cong Lorenz*

* Đường chéo (đường bình đẳng hoàn hảo): Đại diện cho sự phân phối thu nhập hoặc tài sản hoàn toàn bình đẳng.
* Đường cong Lorenz: Biểu diễn phân phối thực tế của thu nhập hoặc tài sản.

#### 2.6.3.4. K-S Statistic

K-S Statistic (Kolmogorov-Smirnov Statistic) là một chỉ số trong kiểm định Kolmogorov-Smirnov, được sử dụng để đánh giá sự khác biệt giữa hai phân phối xác suất (hoặc giữa một mẫu thực nghiệm và một phân phối chuẩn).

K-S Statistic đo lường khoảng cách lớn nhất (maximum distance) giữa các hàm phân phối tích lũy (CDF - Cumulative Distribution Function) của hai phân phối.



*Hình 3.0 Công thức tính K-S Statistic*

### 2.6.4. Phương pháp tìm ngưỡng phân loại

Trong đề tài nghiên cứu này, đầu ra của các mô hình dữ liệu là giá trị xác suất trong khoảng (0,1) chỉ báo xác suất vỡ nợ của khách hàng. Từ giá trị xác suất này cần tìm được một ngưỡng giá trị mà tại đó các giá trị xác suất được chuyển đổi chính xác nhất thành giá trị nhị phân nhằm phân loại khách hàng có rủi ro tín dụng và khách hàng không có rủi ro tín dụng. Tìm kiếm ngưỡng phân loại tốt nhất cho mô hình dự đoán có thể dựa vào một trong số các số chỉ số đánh giá sau: accuracy, precision, recall, F1-Score, sensitivity(TPR), specificity, roc\_auc.

Quá trình tìm ngưỡng phân loại sẽ sử dụng phương pháp thống kê Youden's J [101] với mục tiêu tìm giá trị lớn nhất của 𝓙:

𝓙 = sensitivity + specificity − 1 *PT 2.29*

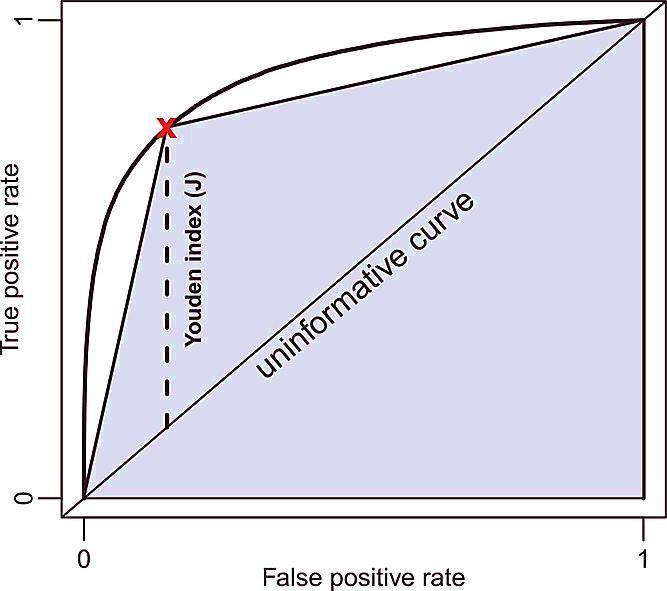
= TPR − (1 − FPR) − 1

= TPR − FPR

Các giá trị sensitivity(TPR), specificity, FPR là các giá trị đã được đề cập trong phần *Phương pháp đánh giá mô hình.*

Các bước tìm ngưỡng phân loại bao gồm:

1. Xây dựng đường cong ROC bằng cách sử dụng xác suất dự đoán của mô hình và nhãn thực tế.
2. Lấy ra các bộ giá trị bao gồm (FPR, TPR, threshold) quá trình xây dựng đường cong ROC.
3. Tính toán chỉ số thống kê 𝓙 cho mỗi bộ giá trị (FPR, TPR, threshold).
4. Với bộ giá trị có chỉ số 𝓙 lớn nhất. Giá trị threshold là giá trị ngưỡng phân loại tốt nhất của mô hình.



*Hình 3.1 Minh họa giá trị ngưỡng phân loại tốt nhất trên đường cong ROC [102]*

### 2.6.5. Phương pháp chuyển đổi điểm tín dụng từ kết quả đầu ra mô hình

Trong chủ đề nghiên cứu này mục tiêu của mô hình chấm điểm tín dụng sẽ là điểm tín dụng với các thuộc tính dữ liệu người dùng ở đầu vào. Tuy nhiên, mô hình dự đoán với đầu ra sẽ là xác xuất trong khoảng (0,1) thể hiện xác suất vỡ của khách hàng. Mục tiêu của phương pháp chuyển đổi điểm tín dụng từ kết quả đầu ra mô hình sẽ chuyển đổi xác suất từ đâu ra của mô hình thành điểm tín dụng trong khoảng (300,850).

Với mỗi mẫu dự đoán sẽ được chuyển đổi với công thức:

𝐂𝐒 = 𝛼 × 300 + (1 − 𝛼) × 850 *PT 2.32*

Trong đó:

* 𝐂𝐒: Điểm tín dụng
* 𝛼: Giá trị dự đoán của mô hình

Ví dụ, nếu giá trị dự đoán của mô hình là 0.5, khi đó điểm tín dụng của khách hàng ứng với mẫu dữ liệu mới sẽ là: 𝐂𝐒 = (0.5 × 300 + (1 − 0.5) × 850) = 575.

Như vậy điểm tín dụng của khách hàng này sẽ là 575 điểm.

# CHƯƠNG 3. DỮ LIỆU

## 3.1. Tổng quan dữ liệu

Bộ dữ liệu “Home Credit Default Risk” được sử dụng rộng rãi trong ngành tài chính, đặc biệt là để đánh giá rủi ro tín dụng và dự đoán vỡ nợ. Bộ dữ liệu được tổng hợp và công khai bởi Home Credit Group, một tổ chức tài chính phi ngân hàng quốc tế chuyên cung cấp các khoản vay cho các cá nhân ít hoặc không có lịch sử tín dụng và chưa được tiếp cận đầy đủ các dịch vụ từ ngân hàng. Bộ dữ liệu được cung cấp công khai trên Kaggle [8], bao gồm nhiều tệp dữ liệu định dạng CSV, mỗi tệp cung cấp thông tin về các khía cạnh khác nhau liên quan đến dữ liệu tín dụng của mỗi khách hàng đăng ký khoản vay.



*Hình 3.2 Logo Home Credit Group*

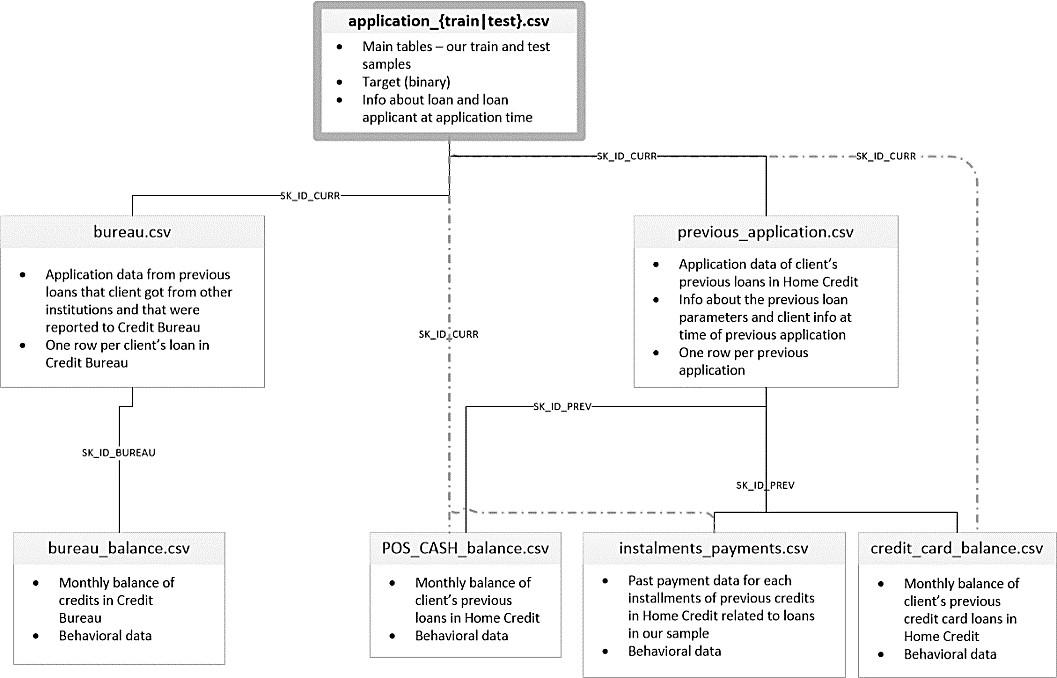
Tệp dữ liệu chính trong bộ dữ liệu là tệp dữ liệu “application\_train. csv”, tệp dữ liệu này chứa thông tin về khách hàng đăng ký khoản vay và biến mục tiêu nhị phân (0 hoặc 1) cho biết liệu rằng người đăng ký khoản vay có gặp khó khăn trong việc thanh toán các khoản vay trước đó đúng hạn hay không với giá trị 0 thể hiện khách hàng đăng ký vay không gặp khó khăn trong việc hoàn trả nợ và 1 thể hiện khách hàng đăng ký vay gặp khó khăn trong việc thanh toán nợ và đây là khách hàng có rủi ro tín dụng cao. Các tệp khác trong tập dữ liệu cung cấp thông tin bổ sung về dữ liệu tín dụng có liên quan đến khách hàng đăng ký khoản vay, các khoản vay đã đăng ký trước đó trước đó, số dư thẻ tín dụng và các thông tin liên quan khác.

Trong bộ dữ liệu này có đề cập thêm thông tin dữ liệu tín dụng từ các tổ chức

Credit Bureau. Đây là những văn phòng tín dụng thu thập thông tin tín dụng từ nhiều tổ chức tài chính và phi tài chính, bao gồm các tổ chức tài chính vi mô và công ty tín dụng, đồng thời cung cấp thông tin tín dụng tiêu dùng toàn diện. Những văn phòng tín dụng này có xu hướng thu thập dữ liệu rất chi tiết về các khách hàng cá nhân, do đó các thông tin thường toàn diện hơn và được thiết kế tốt hơn để đánh giá và giám sát mức độ tín nhiệm của các khách hàng cá nhân [67]. Trong bộ dữ liệu Home Credit Default Risk cũng có sử dụng các thông tin về những khoản tín dụng trước đây của các khách hàng đăng ký khoản vay ở Home Credit.

Bộ dữ liệu chứa nhiều biến số về thông tin khách hàng đăng ký khoản vay như dữ liệu nhân khẩu học, lịch sử tín dụng, thu nhập, lịch sử việc làm và nhiều yếu tố khác có thể được sử dụng để dự đoán mức độ tin cậy của khách hàng đăng ký khoản vay.

## 3.2. Mô tả dữ liệu



*Hình 3.3 Tổng quan mối quan hệ giữa các tệp dữ liệu trong bộ dữ liệu Home Credit Default Risk*

*Bảng 3.1 Thông tin về các tập dữ liệu trong bộ dữ liệu Home Credit Default Risk*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Tên tệp dữ liệu** | **Số lượng hàng** | **Số lượng cột** | **Định dạng** | **Số lượng thuộc tính categorical** | **Số lượng thuộc tính numeric** |
| 1 | application\_train | 307,511 | 122 | CSV | 16 | 106 |
| 2 | application\_test | 48,744 | 121 | CSV | 16 | 105 |
| 3 | bureau | 1,716,428 | 17 | CSV | 3 | 14 |
| 4 | bureau\_balance | 27,299,925 | 3 | CSV | 1 | 2 |
| 5 | previous\_application | 1,670,214 | 37 | CSV | 16 | 21 |
| 6 | POS\_CASH\_balance | 10,001,203 | 8 | CSV | 4 | 4 |
| 7 | credit\_card\_balance | 38,067,841 | 23 | CSV | 4 | 19 |
| 8 | installments\_payments | 13,605,401 | 8 | CSV | 1 | 7 |

Dữ liệu application{train|test}. csv:

* Đây là tập dữ liệu chính đã được chia thành 2 tập dữ liệu dành cho quá trình huấn luyện (có biến mục tiêu “TARGET”) và tập dữ liệu kiểm thử (không có biến mục tiêu “TARGET”)
* Đây là dữ liệu cố định cho các khách hàng đăng ký khoản vay. Mỗi hàng dữ liệu đại diện một khoản vay.

Dữ liệu bureau. csv:

* Dữ liệu về các khoản tín dụng trước đây của các khách hàng đăng ký vay được cung cấp bởi các phòng tín dụng bao gồm thông tin về loại khoản vay, số tiền cho vay, số ngày quá hạn thanh toán, …
* Đối với mỗi khách hàng đăng ký vay sẽ có nhiều hàng thể hiện số lượng tín dụng mà khách hàng đã có trong báo cáo từ phòng tín dụng trước ngày nộp đơn đăng ký vay ở Home Credit.

Dữ liệu bureau\_balance. csv

* Dữ liệu về số dư hàng tháng của khách hàng đăng ký vay tín dụng trước đây và có thông tin trong dữ liệu bureau.csv.
* Với mỗi khoản tín dụng, mỗi hàng trong dữ liệu thể hiện một tháng trong lịch sử của các khoản tín dụng.

Dữ liệu POS\_CASH\_balance. csv:

* Dữ liệu chứa thông tin về số dư hàng tháng của các điểm bán hàng (POS) và các khoản vay tiền mặt mà khách hàng đã đăng ký với Home Credit trước đây.
* Mỗi hàng dữ liệu thể hiện thông tin giao dịch tín dụng mỗi tháng của tất cả

các khoản tín dụng trước đây (tín dụng tiêu dùng và cho vay tiền mặt).

Dữ liệu credit\_card\_balance. csv:

* Dữ liệu chứa dữ liệu số dư hàng tháng của các giao dịch thẻ tín dụng mà khách hàng đã đăng ký với Home Credit trước đây.
* Mỗi hàng dữ liệu thể hiện thông tin giao dịch tín dụng mỗi tháng của tất cả các khoản tín dụng trước đây của khách hàng đã từng đăng ký (tín dụng tiêu dùng và cho vay tiền mặt).

Dữ liệu previous\_application. csv:

* Dữ liệu chứa thông tin về các khoản tín dụng mà khách hàng đã đăng ký trước đây bao gồm mã khách hàng đăng ký, số tiền vay, ngày nộp đơn, trạng thái phê duyệt, …
* Mỗi hàng trong dữ liệu thể hiện một khoản tín dụng mà khách hàng đã đăng ký trước đó.

Dữ liệu installments\_payments. csv:

* Dữ liệu chứa thông tin về lịch sử thanh toán nợ cho các khoản tín dụng đã giải ngân trước đây mà khách hàng đã từng đăng ký với Home Credit.
* Mỗi hàng tương ứng với một lần thanh toán của các khoản tín dụng thanh toán một lần hoặc một lần trả góp cho các khoản tín dụng trả góp.

## 3.3. Chuẩn bị dữ liệu

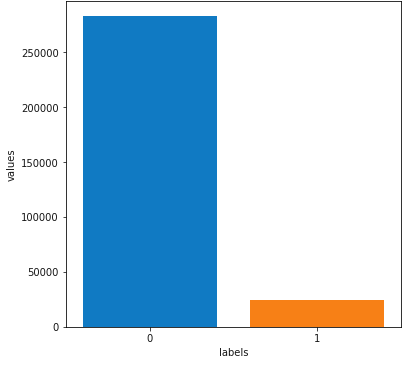
### 3.3.1. Phân tích khám phá dữ liệu

Quá trình phân tích dữ liệu trên bộ dữ liệu Home Credit Default Risk liên quan đến việc tạo các biểu đồ trực quan để khám phá dữ liệu và xác định các mẫu ngoại lệ. Những thông tin trực quan này sẽ giúp xác định xu hướng, mối quan hệ và sự bất thường trong dữ liệu từ đó cung cấp thông tin về bộ dữ liệu, phục vụ cho quá trình xử lý và trích xuất đặc trưng [68].

#### 3.3.1.1. Cân bằng dữ liệu trên biến mục tiêu

Biến mục tiêu trong bộ dữ liệu đào tạo có 2 nhãn (label) với:

* 0: khoản vay đã được hoàn trả.
* 1: khoản vay không được hoàn trả.



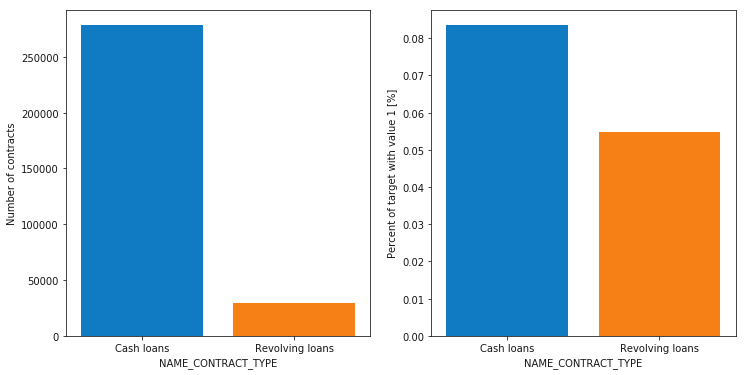
*Hình 3.4 Đồ thị biểu diễn số lượng nhãn phân loại của 2 lớp trong biến mục tiêu*

Biểu đồ cho thấy rằng dữ liệu có sự mất cân bằng rất lớn giữa 2 lớp phân loại.

Vấn đề này có thể dẫn đến việc mô hình có độ chính xác cao đối với lớp 0 nhưng hoạt động kém đối với lớp 1 [69].

#### 3.3.1.2. Loại khoản vay

Tìm hiểu các loại khoản vay được thực hiện và tỷ lệ của các loại khoản vay trên các trường hợp không thể trả nợ (label=1).

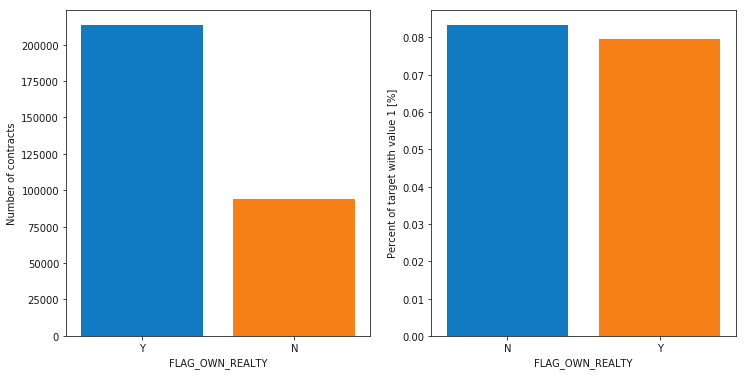


*Hình 3.5 Biểu đồ số lượng loại khoản vay được thực hiện (phải) và tỷ lệ của các loại khoản vay trên các trường hợp không thể trả nợ (trái)*

Loại khoản vay “Revolving loans” chỉ chiếm một phần nhỏ trong tổng số các khoản vay (khoảng 10%), chiếm phần lớn là “Cash loans” (Cho vay tiền mặt). Tuy nhiên các khoản vay “Revolving loans” vẫn chiếm một phần lớn trong các khoản vay không thể thanh toán.

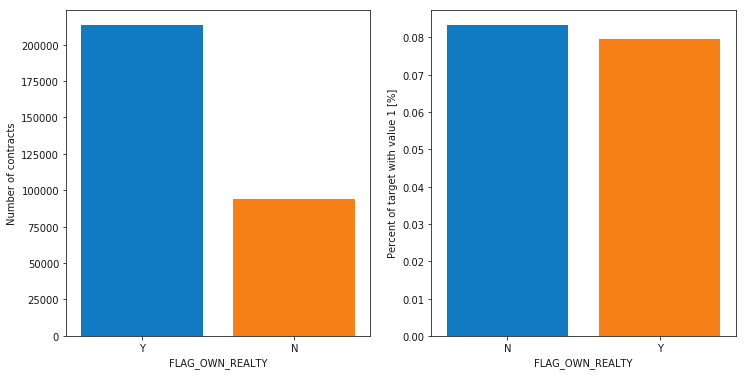
#### 3.3.1.3. Dữ liệu khách hàng có sở hữu ô tô hoặc bất động sản

Tìm hiểu số lượng lượng khách hàng có sở hữu ô tô hoặc bất động sản trong bộ dữ liệu và tỷ lệ trong các khoản vay không thể trả (label=1).



*Hình 3.6 Biểu đồ số lượng khách hàng sở hữu ô tô (phải) và tỷ lệ trên các khoản vay không thể trả (trái)*

Những khách hàng sở hữu ô tô gần gấp đôi số khách hàng không sở hữu ô tô. Tuy nhiên, cả hai loại (sở hữu ô tô hoặc không sở hữu) đều có tỷ lệ không trả nợ khoảng 8%.

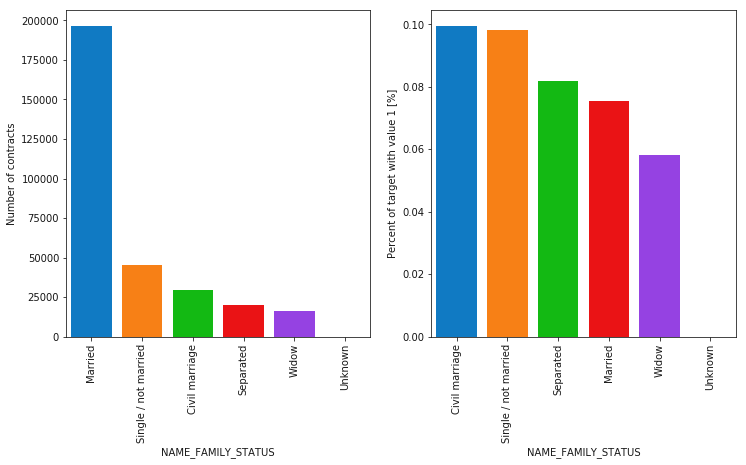


*Hình 3.7 Biểu đồ số lượng khách hàng sở hữu bất động sản (phải) và tỷ lệ trên các khoản vay không thể trả (trái)*

Những khách hàng sở hữu bất động sản nhiều hơn gấp đôi so với những khách hàng không sở hữu. Cả hai loại (sở hữu bất động sản hoặc không sở hữu) đều có tỷ lệ không trả nợ xấp xỉ 8%.

#### 3.3.1.4. Tình trạng gia đình

Tìm hiểu về tình trạng gia đình của các khách hàng vay và tỷ lệ trên các khoản vay không thể thanh toán (label=1).



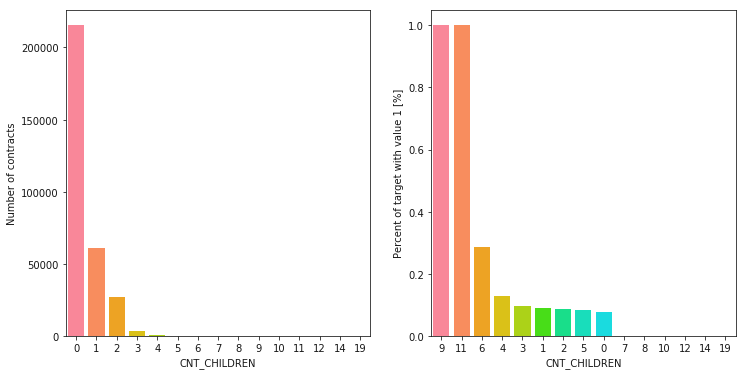
*Hình 3.8 Biểu đồ tình trạng gia đình của các khách hàng vay (phải) và tỷ lệ trên các khoản vay không thể thanh toán (trái)*

Hầu hết các khách hàng đã kết hôn (Married), tiếp theo là độc thân/chưa kết hôn (Single/not married) và hôn nhân dân sự (Civil marriage).

Xét trên tỷ lệ không hoàn trả nợ, hôn nhân dân sự (Civil marriage) có tỷ lệ không trả được nợ cao nhất (10%), khách hàng là góa phụ (Widow) có tỷ lệ thấp nhất.

#### 3.3.1.5. Thành viên gia đình

Tìm hiểu về phân phối số lượng thành viên trong gia đình của các khách hàng vay và tỷ lệ trên các khoản vay không thể thanh toán (label=1).

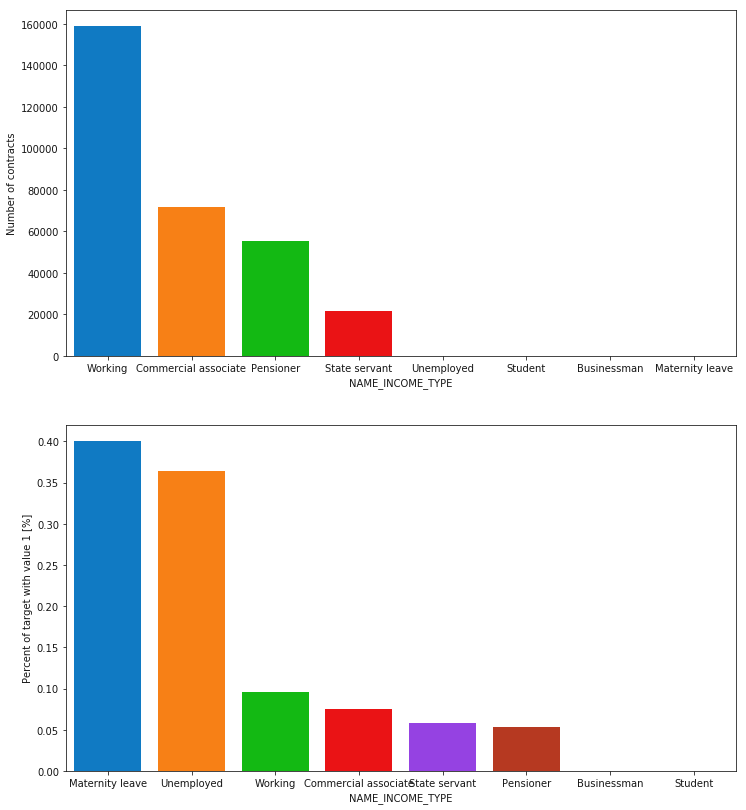


*Hình 3.9 Biểu đồ phân phối số lượng thành viên trong gia đình (phải) của các khách hàng vay và tỷ lệ trên các khoản vay không thể thanh toán (trái)*

Khách hàng có thành viên gia đình gồm 2 người có số lượng khoản vay nhiều nhất, tiếp theo là gia đình có 1 người (độc thân), 3 người và 4 người.

Các gia đình có 10 hoặc 8 thành viên có tỷ lệ không hoàn trả nợ trên 30%. Các gia đình có từ 6 thành viên trở xuống có tỷ lệ không hoàn trả nợ ở mức trung bình 10%. Khách hàng có số lượng người trong gia đình từ 11 đến 13 người có tỷ lệ không hoàn trả nợ là 100% (Đây có thể là dữ liệu sai lệch hoặc ngoại lai cần xử lý).

#### 3.3.1.6. Loại thu nhập



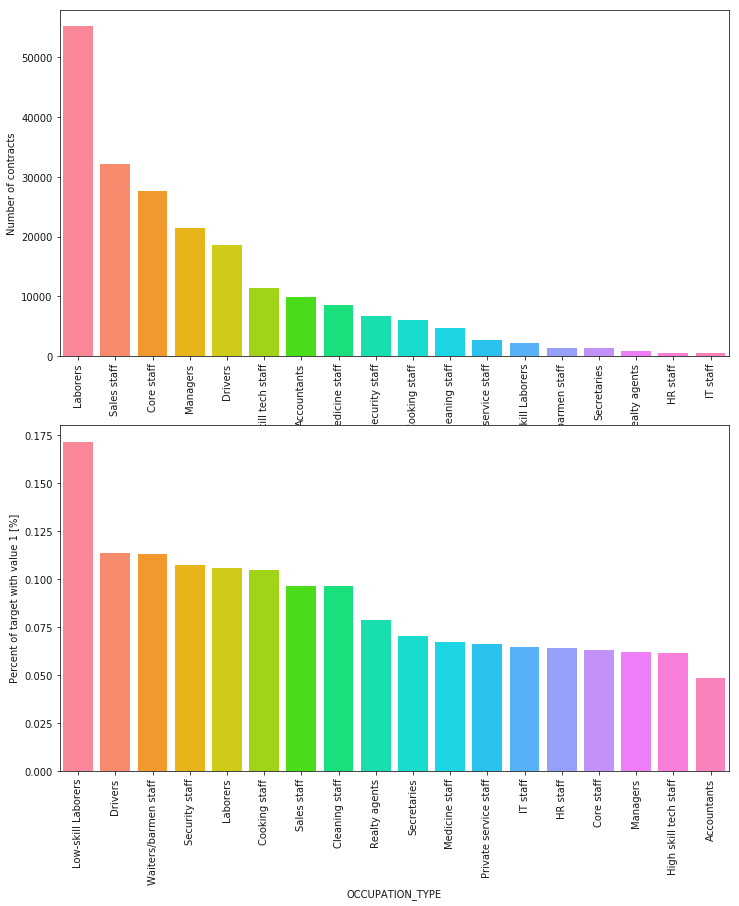
*Hình 3.10 Biểu đồ thể hiện số lượng khách hàng vay với các loại thu nhập khác nhau và tỷ lệ trên các khoản vay không hoàn trả*

Hầu hết những khách hàng đăng ký vay đều có thu nhập từ làm việc (Working), tiếp theo là cộng tác viên thương mại (Commercial associate), hưu trí (Pensioner) và công chức nhà nước (State servant).

Các khách hàng có thu nhập từ trợ cấp nghỉ thai sản (Maternity leave) có tỷ lệ không trả nợ cao nhất (gần 40%), tiếp theo là thất nghiệp (Unemployed) với tỷ lệ

37%. Các loại thu nhập còn lại dưới mức trung bình 10% đối với các khoản vay không hoàn trả nợ.

#### 3.3.1.7. Ngành nghề làm việc

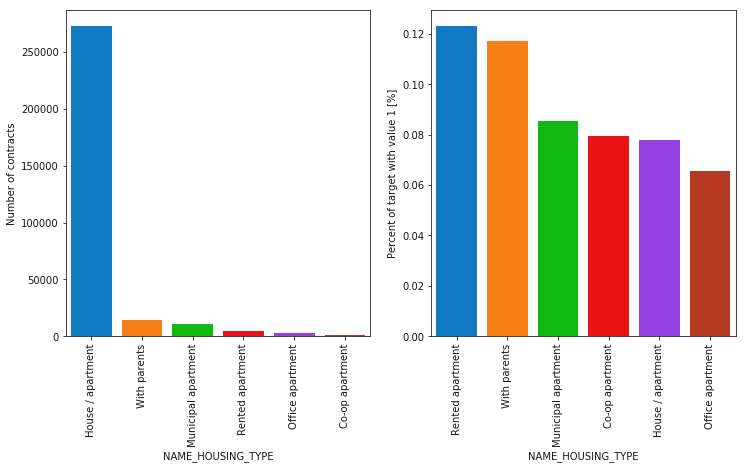


*Hình 3.11 Biểu đồ phân phối ngành nghề làm việc của khách hàng vay và tỷ lệ trên số lượng không hoàn trả nợ*

Hầu hết các khoản vay được vay bởi người lao động (Laborers), tiếp theo là nhân viên kinh doanh (Sales staff). Nhân viên công nghệ thông tin có số lượng đăng ký vay thấp nhất (IT staff).

Nhóm có tỷ lệ không hoàn trả nợ cao nhất là lao động kỹ năng thấp (Low-skill Laborers) với trên 17%, tiếp theo là nhân viên lái xe (Drivers) và nhân viên phục vụ (Waiters/barmen staff), nhân viên bảo vệ (Security staff), người lao động (Laborers) và Nhân viên nấu ăn (Cooking staff).

#### 3.3.1.8. Loại nhà ở và cư trú



*Hình 3.12 Biểu đồ phân phối số lượng của các loại nhà ở/cư trú của các khách hàng vay và tỷ lệ trên số lượng không hoàn trả nợ*

Hơn 250.000 khách hàng đăng ký vay tín dụng với loại nhà ở là nhà/căn hộ

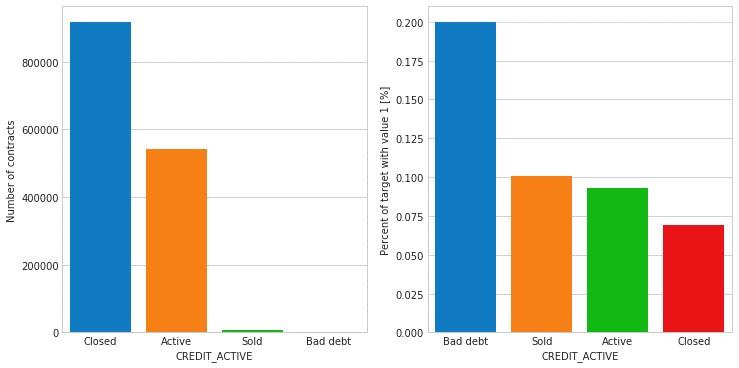
(House/apartment). Các danh mục sau có số lượng khách hàng rất nhỏ (Ở với cha mẹ (With parents), căn hộ thành phố (Municipal apartment).

Các khách hàng ở căn hộ cho thuê (Rented apartment) và ở cùng cha mẹ (With parents) có tỷ lệ không trả nợ cao hơn 10%.

#### 3.3.1.9. Tình trạng tín dụng

Tìm hiểu sự phân bố tình trạng tín dụng với:

* Closed: Tín dụng đã đóng.
* Active: Tín dụng đang hoạt động.
* Sold: Tín dụng đã được bán.
* Bad debt: Tín dụng xấu.



*Hình 3.13 Phân phối loại tín dụng và tỷ lệ trên các khoản vay không hoàn trả nợ*

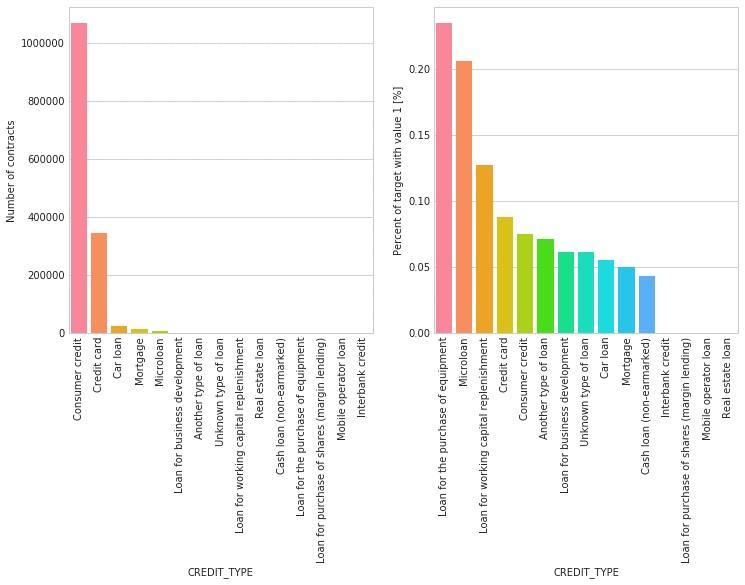
Hầu hết các khoản tín dụng đã đăng ký tại phòng tín dụng đều ở trạng thái “Closed”

(Khoản 900 nghìn khoản tín dụng). Tiếp theo là các khoản tín dụng “Active” (khoảng 600 ngàn khoản tín dụng). Các khoản tín dụng có trạng thái “Sold” và “Bad debt” chiếm số lượng ít.

Các khách hàng có tín dụng đã đăng ký với phòng tín dụng có trạng thái “Bad debt” có tỷ lệ không hoàn trả nợ cao (Khoảng 20%). Khách hàng có trạng thái tín dụng “Sold”, “Active” và “Closed” có tỷ lệ không hoàn trả nợ nhỏ hơn 10%.

Như vậy, lịch sử tín dụng đã đăng ký trước đây là một yếu tố dự đoán mạnh mẽ cho các khoản tín dụng sai phạm.

#### 3.3.1.10. Loại tín dụng



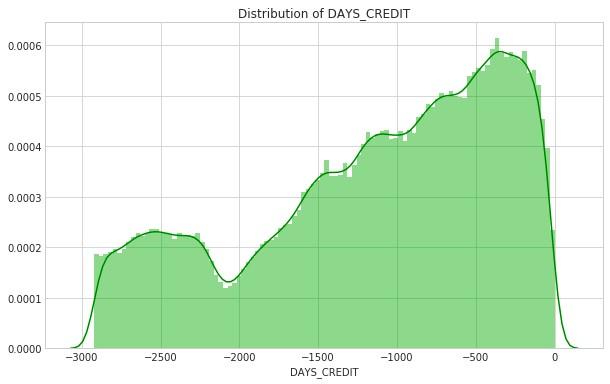
*Hình 3.14 Biểu đồ phân phối của các loại tín dụng khác nhau và tỷ lệ trên các khoản vay không hoàn trả nợ*

Phần lớn các khoản tín dụng được đăng ký tại phòng tín dụng là “Consumer credit” (tín dụng tiêu dùng) và “Credit card” (thẻ tín dụng). Số lượng tín dụng chiếm số lượng nhỏ hơn là “Car loan” (Tín dụng mua/thuê ô tô), “Mortgage” (Tín dụng thế chấp tài sản) và “Microloan” (Tín dụng với các khoản vay rất nhỏ).

Với các loại tín dụng đã đăng ký tại phòng tín dụng trước đây, có một số loại có tỷ lệ vỡ nợ tín dụng cao như:

* “Loan for the purchase of equipment” (Khoản vay mua thiết bị) và “Microloan”, mỗi loại chiếm hơn 20% số lượng các khoản vay không hoàn trả nợ.
* “Loan for working capital replenishment” (Khoản vay bổ sung vốn lưu động) với hơn 12% trên các khoản nợ không hoàn trả nợ.

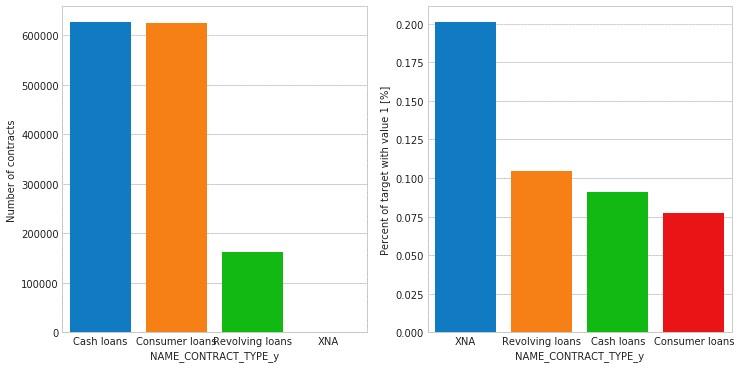
#### 3.3.1.11. Thời hạn tín dụng



*Hình 3.15 Phân phối thời hạn tín dụng của các khoản tín dụng đã vay trước đây từ dữ liệu của phòng tín dụng*

Thời hạn tín dụng được tính theo ngày dao động từ 500 ngày đến 3000 ngày. Từ biểu đồ có thể thấy phân phối các các khoản vay nhiều nhất ở khoảng từ 200 đến 1000 ngày và mức cao nhất ở khoảng 300 ngày. Các khoản vay thường được đăng ký với thời hạn tín dụng từ 200 đến 100 ngày.

#### 3.3.1.12. Loại hợp đồng tín dụng



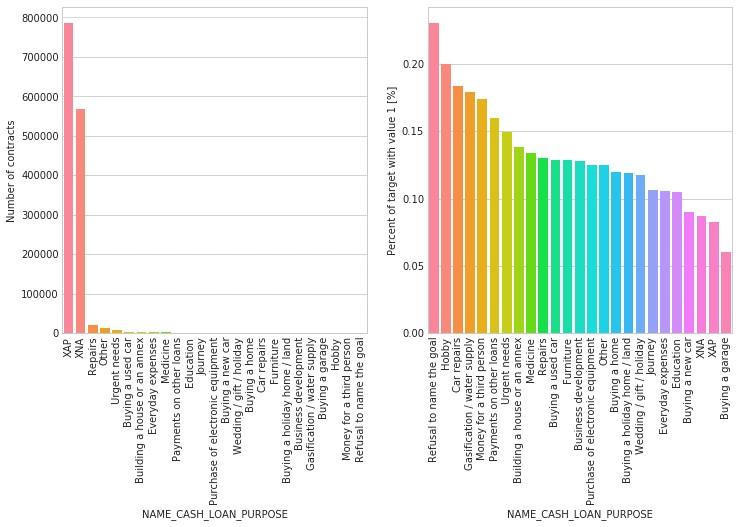
*Hình 3.16 Phân bố số lượng của các loại hợp đồng tín dụng khác nhau và tỷ lệ trên các khoản vay không hoàn trả nợ*

Có 3 loại hợp đồng trong dữ liệu đăng ký tín dụng trước đây bao gồm: “Cash loans” (Vay tiền mặt), “Consumer loans” (vay tiêu dùng), “Revolving loans” (Vay quay vòng). Số lượng khoản vay tiền mặt và vay tiêu dùng gần bằng nhau (Khoảng 600 nghìn khoản vay) trong khi số lượng khoản vay quay vong khá thấp (150 nghìn khoản vay).

Tỷ lệ các khoản vay không hoàn trả nợ đối với các khách hàng đăng ký khoản vay trước đó khác nhau đối với loại hợp đồng, giảm từ 10% đối với khoản vay quay vòng, 9,5% đối với khoản vay tiền mặt và 8% đối với khoản vay tiêu dùng.

#### 3.3.1.13. Mục đích vay tiền mặt

Tìm hiểu về mục đích vay tiền mặt đối với các khoản vay tiền mặt.



*Hình 3.17 Phân bố số lượng của các mục đích vay tiền mặt khác nhau và tỷ lệ trên số lượng khoản vay không hoàn trả nợ*

Trừ các khoản vay không xác định (XAP và XNA), các khoản vay “Repairs” (khoản vay dùng để sửa chữa), “Urgent needs” (khoản vay dùng cho nhu cầu cấp thiết), “Buying a used car” (Khoản vay dùng để mua xe cũ) và “Building a house or an annex” (khoản vay dùng để xây nhà hoặc công trình phụ) chiếm số lượng lớn các khoản vay tín dụng.

Tỷ lệ các khoản vay không hoàn trả nợ cao nhất ở khoản vay “Refusal to name the goal” (Khoản vay không có lý do) chiếm 23%, “Hobby” (khoản vay phục vụ sở thích) chiếm (20%) và “Car repairs” (khoản vay để sử xe) chiếm 18%.

Từ các biểu đồ và phân tích [70] thông qua các thuộc tính trong bộ dữ liệu Home Credit Default Risk có thể nắm bắt được các thuộc tính quan trọng ảnh hưởng đến tỷ lệ không hoàn trả nợ vay. Từ quá trình này, tiếp tục thực hiện trích xuất các tính năng từ những thông tin đã nắm bắt.

### 3.3.2. Trích xuất đặc trưng dữ liệu

Sau khi hoàn thành các bước phân tích và khám phá dữ liệu ban đầu, bước quan trọng tiếp theo là trích xuất tính năng. Trong bước này, từ thuộc tính có liên quan nhất đến các các yếu tố thanh toán khoản vay trong bộ dữ liệu tiến hành tạo thêm các thuộc tính mới bằng cách chọn các tính năng mang tính dự báo, cung cấp nhiều thông tin và sử dụng các kiến thức có liên quan đến tài chính tín dụng [71], từ đó có thể khai thác đầy đủ các giá trị ẩn trong dữ liệu phục vụ quá trình xây dựng các mô hình [72].

**3.3.2.1. Tập dữ liệu** 𝐏𝐎𝐒\_𝐂𝐀𝐒𝐇\_𝐛𝐚𝐥𝐚𝐧𝐜𝐞Sử dụng các hàm tổng hợp để lấy ra:

* 𝐏𝐎𝐒\_𝐂𝐍𝐓\_𝐈𝐍𝐒𝐓𝐀𝐋𝐌𝐄𝐍𝐓\_𝐅𝐔𝐓𝐔𝐑𝐄\_𝐌𝐈𝐍: Số tiền tối thiểu còn lại để thanh toán cho khoản tín dụng.
* 𝐏𝐎𝐒\_𝐍𝐀𝐌𝐄\_𝐂𝐎𝐍𝐓𝐑𝐀𝐂𝐓\_𝐒𝐓𝐀𝐓𝐔𝐒\_𝐂𝐨𝐦𝐩𝐥𝐞𝐭𝐞𝐝\_𝐒𝐔𝐌: Số lượng trạng thái hợp đồng là “complete”*.*
* Nếu đồng thời𝐏𝐎𝐒\_𝐂𝐍𝐓\_𝐈𝐍𝐒𝐓𝐀𝐋𝐌𝐄𝐍𝐓\_𝐅𝐔𝐓𝐔𝐑𝐄\_𝐌𝐈𝐍và

𝐏𝐎𝐒\_𝐍𝐀𝐌𝐄\_𝐂𝐎𝐍𝐓𝐑𝐀𝐂𝐓\_𝐒𝐓𝐀𝐓𝐔𝐒\_𝐂𝐨𝐦𝐩𝐥𝐞𝐭𝐞𝐝\_𝐒𝐔𝐌đều bằng 0 thì sẽ

đánh nhãn cho các khoản tín dụng này không trả đúng hạn

𝐏𝐎𝐒\_𝐍𝐄𝐖\_𝐈𝐒\_𝐂𝐑𝐄𝐃𝐈𝐓\_𝐍𝐎𝐓\_𝐂𝐎𝐌𝐏𝐋𝐄𝐓𝐄𝐃\_𝐎𝐍\_𝐓𝐈𝐌𝐄sẽ là:

* + 1: Khoản vay chưa được đóng đúng hạn.
  + 0: Khoản vay đã được đóng đúng hạn.

Các thuộc tính còn lại của tập dữ liệu 𝐏𝐎𝐒\_𝐂𝐀𝐒𝐇\_𝐛𝐚𝐥𝐚𝐧𝐜𝐞sẽ sử dụng các hàm aggregate (min, max, mean, count, sum) để tạo ra các thuộc tính mới.

**3.3.2.2. Tập dữ liệu** 𝐜𝐫𝐞𝐝𝐢𝐭\_𝐜𝐚𝐫𝐝\_𝐛𝐚𝐥𝐚𝐧𝐜𝐞

Tạo thuộc tính 𝐒𝐊\_𝐃𝐏𝐃: Số ngày mà khoản vay đó bị trễ hạn (day past due).

Tính số lần xảy ra các khoản thanh toán chậm dựa trên thuộc tính SK\_DPD

(Giá trị khác 0).

Tạo thuộc tính 𝐀𝐌𝐓\_𝐈𝐍𝐒𝐓\_𝐌𝐈𝐍\_𝐑𝐄𝐆𝐔𝐋𝐀𝐑𝐈𝐓𝐘: Khoản thanh toán tối thiểu cho khoản tín dụng vào tháng đó.

Tạo thuộc tính 𝐀𝐌𝐓\_𝐏𝐀𝐘𝐌𝐄𝐍𝐓\_𝐂𝐔𝐑𝐑𝐄𝐍𝐓: Số tiền mà khách hàng đã thanh toán cho khoản tín dụng của tháng đó.

Với mỗi khách hàng tính tổng số lượng giao dịch có khoảng thanh toán nhỏ hơn khoản thanh toán tối thiểu mà khoản tín dụng đó đặt ra

(𝐀𝐌𝐓\_𝐏𝐀𝐘𝐌𝐄𝐍𝐓\_𝐂𝐔𝐑𝐑𝐄𝐍𝐓 < 𝐀𝐌𝐓\_𝐈𝐍𝐒𝐓\_𝐌𝐈𝐍\_𝐑𝐄𝐆𝐔𝐋𝐀𝐑𝐈𝐓)**.**

Tạo thuộc tính 𝐏𝐄𝐑𝐂𝐄𝐍𝐓𝐀𝐆𝐄\_𝐌𝐈𝐍\_𝐌𝐈𝐒𝐒𝐄𝐃\_𝐏𝐀𝐘𝐌𝐄𝐍𝐓𝐒 (tỷ lệ phần trăm các khoản thanh toán nhỏ hơn khoản thanh toán tối thiểu):

∑(𝐀𝐌𝐓\_𝐏𝐀𝐘𝐌𝐄𝐍𝐓\_𝐂𝐔𝐑𝐑𝐄𝐍𝐓 < 𝐀𝐌𝐓\_𝐈𝐍𝐒𝐓\_𝐌𝐈𝐍\_𝐑𝐄𝐆𝐔𝐋𝐀𝐑𝐈𝐓)

number of instalments (số lần trả góp)

Tạo thuộc tính 𝐀𝐌𝐓\_𝐃𝐑𝐀𝐖𝐈𝐍𝐆𝐒\_𝐀𝐓𝐌\_𝐂𝐔𝐑𝐑𝐄𝐍𝐓: Số tiền rút trong tháng tại ATM. Tính tổng cho mỗi SK\_ID\_CURR.

Tạo thuộc tính 𝐀𝐌𝐓\_𝐃𝐑𝐀𝐖𝐈𝐍𝐆𝐒\_𝐂𝐔𝐑𝐑𝐄𝐍𝐓: Số tiền rút ra trong tháng. Tính tổng cho mỗi SK\_ID\_CURR. Tạo thuộc tính 𝐂𝐀𝐒𝐇\_𝐂𝐀𝐑𝐃\_𝐑𝐀𝐓𝐈𝐎 = AMT\_DRAWINGS\_ATM\_CURRENT

AMT\_DRAWINGS\_CURRENT

Các thuộc tính còn lại của tập dữ liệu 𝐜𝐫𝐞𝐝𝐢𝐭\_𝐜𝐚𝐫𝐝\_𝐛𝐚𝐥𝐚𝐧𝐜𝐞 sẽ sử dụng các hàm aggregate (min, max, mean, count, sum) để tạo ra các thuộc tính mới.

**3.3.2.3. Tập dữ liệu** 𝐩𝐫𝐞𝐯𝐢𝐨𝐮𝐬\_𝐚𝐩𝐩𝐥𝐢𝐜𝐚𝐭𝐢𝐨𝐧

Tạo thuộc tính 𝐍𝐄𝐖\_𝐋𝐎𝐀𝐍\_𝐑𝐀𝐓𝐄 = Tỷ lệ khoản vay yêu cầu (AMT\_APPLICATION)Khoản tín dụng được cấp (AMT\_CREDIT) Từ các thuộc tính:

* 𝐀𝐌𝐓\_𝐀𝐍𝐍𝐔𝐈𝐓𝐘: Khoản thanh toán cố định hằng năm (niên kim).
* 𝐂𝐍𝐓\_𝐏𝐀𝐘𝐌𝐄𝐍𝐓: Kỳ hạn của khoản tín dụng.
* 𝐀𝐌𝐓\_𝐂𝐑𝐄𝐃𝐈𝐓: Khoản tiền tín dụng được cấp.

Tạo thuộc tính 𝐈𝐍𝐓𝐄𝐑𝐄𝐒𝐓\_𝐑𝐀𝐓𝐄:

12

= ((𝐀𝐌𝐓\_𝐀𝐍𝐍𝐔𝐈𝐓𝐘 × 𝐂𝐍𝐓\_𝐏𝐀𝐘𝐌𝐄𝐍𝐓)/(𝐀𝐌𝐓\_𝐂𝐑𝐄𝐃𝐈𝐓 ))𝐂𝐍𝐓\_𝐏𝐀𝐘𝐌𝐄𝐍𝐓 − 1

Các thuộc tính còn lại của tập dữ liệu 𝐜𝐫𝐞𝐝𝐢𝐭\_𝐜𝐚𝐫𝐝\_𝐛𝐚𝐥𝐚𝐧𝐜𝐞sẽ sử dụng các hàm aggregate (min, max, mean, count, sum) để tạo ra các thuộc tính mới.

**3.3.2.4. Tập dữ liệu** 𝐢𝐧𝐬𝐭𝐚𝐥𝐥𝐦𝐞𝐧𝐭𝐬\_𝐩𝐚𝐲𝐦𝐞𝐧𝐭𝐬

Từ 𝐃𝐀𝐘𝐒\_𝐈𝐍𝐒𝐓𝐀𝐋𝐌𝐄𝐍𝐓 (Thời gian của đợt trả góp của khoản tín dụng đó tính đến ngày application ) và 𝐃𝐀𝐘𝐒\_𝐄𝐍𝐓𝐑𝐘\_𝐏𝐀𝐘𝐌𝐄𝐍𝐓 (Thời gian từ lúc chi trả đợt trả góp đó đến ngày application). Nếu khoảng thời gian:

𝐃𝐀𝐘𝐒\_𝐈𝐍𝐒𝐓𝐀𝐋𝐌𝐄𝐍𝐓 − 𝐃𝐀𝐘𝐒\_𝐄𝐍𝐓𝐑𝐘\_𝐏𝐀𝐘𝐌𝐄𝐍𝐓 > 𝟎 gán nhãn (label) là 0

(trả sớm). Ngược lại gán nhãn 1 (trả muộn)

Ví dụ: 𝐃𝐀𝐘𝐒\_𝐈𝐍𝐒𝐓𝐀𝐋𝐌𝐄𝐍𝐓 = −1180

𝐃𝐀𝐘𝐒\_𝐄𝐍𝐓𝐑𝐘\_𝐏𝐀𝐘𝐌𝐄𝐍𝐓 = −1187

⇨ Đợt trả góp này là trả sớm (label: 0)

Thêm các thuộc tính về tỷ lệ thanh toán khoản vay và dư nợ còn lại của mỗi khoản vay trước đó:

* Tỷ lệ thanh toán khoản vay = ∑(𝐀𝐌𝐓\_𝐏𝐀𝐘𝐌𝐄𝐍𝐓 (Số tiền trả cho từng đợt trả góp))

∑(𝐀𝐌𝐓\_𝐈𝐍𝐒𝐓𝐀𝐋𝐌𝐄𝐍𝐓 (Số tiền phải trả cho từng đợt trả góp))

* Dư nợ = Tổng(𝐀𝐌𝐓\_𝐈𝐍𝐒𝐓𝐀𝐋𝐌𝐄𝐍𝐓) − Tổng(𝐀𝐌𝐓\_𝐏𝐀𝐘𝐌𝐄𝐍𝐓)

Các thuộc tính còn lại của tập dữ liệu 𝐢𝐧𝐬𝐭𝐚𝐥𝐥𝐦𝐞𝐧𝐭𝐬\_𝐩𝐚𝐲𝐦𝐞𝐧𝐭𝐬 sẽ sử dụng các hàm aggregate (min, max, mean, count, sum) để tạo ra các thuộc tính mới.

Sau quá trình trích xuất đặc trưng từ các thuộc tính cơ bản sẽ có bộ dữ liệu mới có thêm các thuộc tính đã trích xuất, các thuộc tính đã được chuyển đổi từ thuộc tính cũ. Với mỗi thuộc tính mới được trích xuất sẽ sử dụng chỉ số đánh giá giá trị dữ liệu (Information value - IV) để đánh giá lượng thông tin mà một thuộc tính cung cấp về biến mục tiêu trong tập dữ liệu nếu cho ra kết quả đủ tốt sẽ đưa vào tạo thành dữ liệu mới và chuyển qua quá trình tổng hợp dữ liệu.

### 3.3.3. Tổng hợp dữ liệu

Dữ liệu sau quá trình trích xuất đặc trưng từ các tập dữ liệu nhỏ sẽ được tổng hợp thành một bộ dữ liệu lớn cho quá trình đào tạo thông qua phương thức hợp nhất

(merge) dựa trên thuộc tính SK\_ID\_CURR (Mã khách hàng đăng ký khoản vay). Sau khi hợp nhất mỗi SK\_ID\_CURR sẽ có các thông tin đầy đủ từ dữ liệu cơ bản và các thuộc tính trong quá trình trích xuất đặc trưng.

Dữ liệu này sẽ là nguồn dữ liệu chính cho quá trình đào tạo các mô hình.

# CHƯƠNG 4. THỰC NGHIỆM VÀ KẾT QUẢ

## 4.1. Thực nghiệm

### 4.1.1. Dữ liệu

Dữ liệu đã được kết hợp từ quá trình chuẩn bị dữ liệu được chia làm 2 phần là dữ liệu huấn luyện (train set) và dữ liệu đánh giá (test set) bao gồm:

* Train set: 356251 mẫu và 781 thuộc tính.
* Test set: 48744 mẫu và 781 thuộc tính.

Train set được sử dụng cho quá trình huấn luyện các mô hình, Test set sẽ sử dụng kết quả dự đoán của mô hình sau khi huấn luyện để dự đoán giá trị xác suất của biến mục tiêu cho các mẫu có trong tập dữ liệu sau đó nộp và đánh giá bằng chỉ số đánh giá của cuộc thi trên nền tảng Kaggle.

### 4.1.2. Huấn luyện mô hình

#### 4.1.2.1. Mô hình logistic regression 4.1.2.1.1. Tham số huấn luyện

*Bảng 4.1 Tham số mô hình Logistic regression*

|  |  |  |
| --- | --- | --- |
| **Tham số** | **Chú thích** | **Giá trị** |
| aggregationDepth | Độ sâu được đề xuất cho tre Aggregate (>= 2). | 2 |
| elasticNetParam | Tham số trộn Elastic Net, trong khoảng [0, 1]. 0: L2 penalty, 1: L1 penalty. | 0.0 |
| family | Loại phân phối nhãn: binomial, multinomial, auto. | auto |
| featuresCol | Tên cột chứa vector đặc trưng (features). | features |
| fitIntercept | Có tính toán hệ số chặn (intercept) hay không. | True |
| maxIter | Số vòng lặp tối đa (>= 0). | 100 |
| regParam | Tham số điều chỉnh (regularization parameter) (>= 0). | 0.0 |
| standardization | Chuẩn hóa dữ liệu đặc trưng trước khi huấn luyện. | True |
| tol | Ngưỡng hội tụ cho thuật toán lặp (>= 0). | 1e-06 |
| threshold | Ngưỡng trong phân loại nhị phân, thuộc phạm vi [0, 1]. | 0.5 |
| probabilityCol | Tên cột chứa xác suất dự đoán. | probability |
| rawPredictionCol | Tên cột chứa giá trị dự đoán thô. | rawPrediction |
| maxBlockSizeInMB | Kích thước bộ nhớ tối đa (MB) cho stacking dữ liệu. | 0.0 |

4.1.2.1.2. Công cụ huấn luyện

Mô hình được tạo bằng ngôn ngữ lập trình python trên Data Science Lab

Cấu hình phần cứng:

* CPU: Intel(R) Xeon(R) Silver 4310 CPU @ 2.10GHz
* RAM: 62 GB
* ROM: 195 GB

#### 4.1.2.2. Mô hình random forest 4.1.2.2.1. Tham số huấn luyện

*Bảng 4.2 Tham số mô hình Random forest*

|  |  |  |
| --- | --- | --- |
| **Tham số** | **Chú thích** | **Giá trị** |
| bootstrap | Sử dụng mẫu bootstrap khi xây dựng cây. | True |
| cacheNodeIds | Nếu False, thuật toán sẽ truyền cây đến các executor để khớp instance với node. Nếu True, thuật toán sẽ cache node ID cho từng instance. | False |
| checkpointInterval | Thiết lập khoảng thời gian checkpoint (>= 1) hoặc tắt checkpoint (-1). Nếu không có thư mục checkpoint được thiết lập trong SparkContext, thiết lập này sẽ bị bỏ qua. | 10 |
| featureSubsetStrategy | Số lượng đặc trưng (features) được sử dụng để chia tại mỗi node. | auto |
| featuresCol | Tên cột chứa đặc trưng (features). | features |
| impurity | Tiêu chí được sử dụng để tính toán độ lợi thông tin (information gain). Các tùy chọn: entropy, gini. | gini |
| labelCol | Tên cột chứa nhãn (label). | label |
| maxBins | Số lượng bins tối đa để phân loại các đặc trưng liên tục. | 32 |
| maxDepth | Độ sâu tối đa của cây. | 5 |
| maxMemoryInMB | Bộ nhớ tối đa (MB) phân bổ cho việc tổng hợp histogram. | 256 |
| minInfoGain | Lợi ích thông tin tối thiểu để một phép chia (split) được chấp nhận. | 0.0 |
| minInstancesPerNode | Số lượng instance tối thiểu trong mỗi node con sau khi chia. | 1 |
| minWeightFractionPerNode | Tỉ lệ tối thiểu của tổng trọng số mẫu trong mỗi node con sau khi chia. | 0.0 |

4.1.2.2.2. Công cụ huấn luyện

Mô hình được tạo bằng ngôn ngữ lập trình python trên Data Science Lab

Cấu hình phần cứng:

* CPU: Intel(R) Xeon(R) Silver 4310 CPU @ 2.10GHz
* RAM: 62 GB
* ROM: 195 GB

#### 4.1.2.3. Mô hình XGBoost

4.1.2.3.1. Tham số huấn luyện

*Bảng 4.3 Tham số huấn luyện mô hình XGBoost*

|  |  |  |
| --- | --- | --- |
| **Tham số** | **Chú thích** | **Giá trị** |
| booster | Loại booster (gồm gbtree, gblinear, dart). | gbtree |
| colsample\_bytree | Tỷ lệ mẫu cho mỗi cây (>= 0 và <= 1). | 1 |
| colsample\_bylevel | Tỷ lệ mẫu cho mỗi cấp độ của cây (>= 0 và <= 1). | 1 |
| colsample\_bynode | Tỷ lệ mẫu cho mỗi nút của cây (>= 0 và <= 1). | 1 |
| gamma | Tham số điều chỉnh (regularization) cho độ phức tạp của mô hình. | 0 |
| max\_depth | Độ sâu tối đa của cây (>= 0). | 6 |
| min\_child\_weight | Trọng số tối thiểu cho mỗi node (>= 0). | 1 |
| max\_delta\_step | Bước thay đổi tối đa khi điều chỉnh độ chính xác mô hình. | 0 |
| subsample | Tỷ lệ mẫu được chọn ngẫu nhiên cho mỗi cây (>= 0 và <= 1). | 1 |
| learning\_rate | Tốc độ học (learning rate), giá trị trong phạm vi [0, 1]. | 0.3 |
| n\_estimators | Số lượng ước lượng cây (estimators). | 100 |
| objective | Loại hàm mục tiêu (objective function), ví dụ: 'reg:squared error' cho hồi quy. | reg:squared error |
| eval\_metric | Hàm đánh giá sử dụng trong quá trình huấn luyện (ví dụ: 'rmse'). | rmse |
| scale\_pos\_weight | Tỷ lệ điều chỉnh cho lớp dương (positive class weight). | 1 |
| seed | Giới hạn cho hạt giống ngẫu nhiên. | 0 |
| alpha | Hệ số L1 regularization (alpha). | 1 |
| lambda | Hệ số L2 regularization (lambda). | 1 |
| tree\_method | Phương pháp xây dựng cây (gồm 'auto', 'exact', 'hist', 'gpu\_hist'). | auto |
| predictor | Loại bộ dự đoán được sử dụng ('auto', 'cpu\_predictor', 'gpu\_predictor'). | auto |
| n\_jobs | Số lượng luồng công việc (threads) trong quá trình huấn luyện. | 1 |
| random\_state | Hạt giống ngẫu nhiên cho phân chia dữ liệu. | 42 |

4.1.2.3.2. Công cụ huấn luyện

Mô hình được tạo bằng ngôn ngữ lập trình python trên Data Science Lab

Cấu hình phần cứng:

* CPU: Intel(R) Xeon(R) Silver 4310 CPU @ 2.10GHz
* RAM: 62 GB
* ROM: 195 GB

#### 4.1.2.7. Huấn luyện

4.1.2.7.1. Phương pháp huấn luyện

Phương pháp huấn luyện bằng chia dữ liệu được phân chia thành hai tập chính: tập huấn luyện và tập kiểm tra. Tập huấn luyện được sử dụng để xây dựng mô hình, giúp mô hình học và tối ưu hóa các tham số. Tập kiểm tra được dùng để đánh giá hiệu suất của mô hình sau khi huấn luyện, nhằm kiểm tra khả năng tổng quát của mô hình đối với dữ liệu chưa thấy.

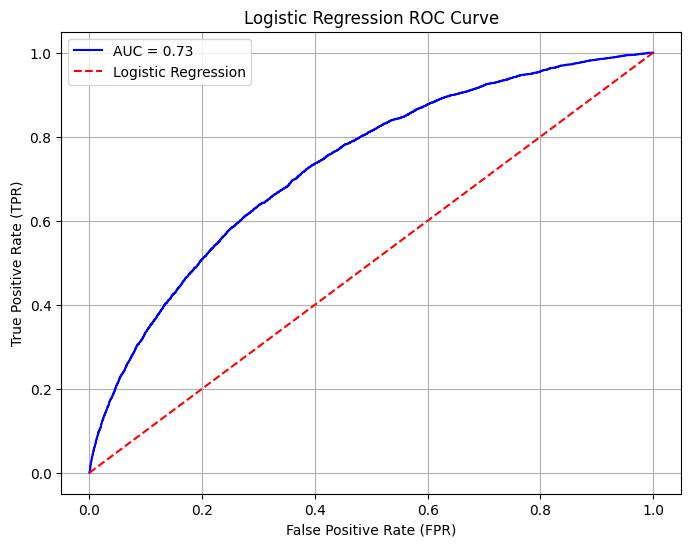
## 4.2. Kết quả

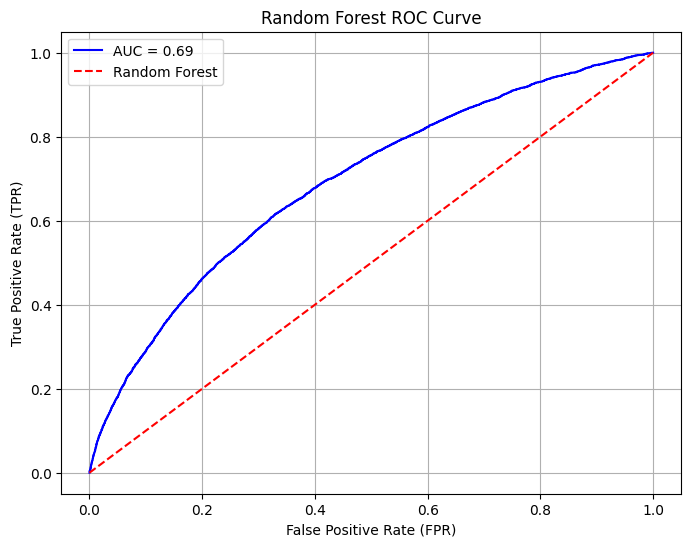
### 4.2.1. Kết quả huấn luyện

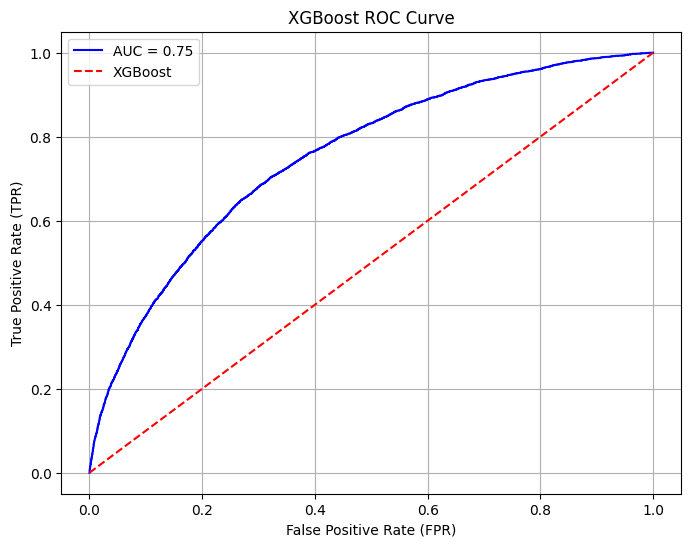
Các chỉ số đánh giá của quá trình huấn luyện này được đề cập dưới đây:

*Bảng 4.4 Chỉ số đánh giá trên biến phân loại*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mô hình** | **Biến phân loại** | **Precision** | **Recall** | **F1-score** |
| Logistic regression | 0 | 0.92 | 0.82 | 0.88 |
| 1 | 0.24 | 0.32 | 0.31 |
| Random forest | 0 | 0.92 | 0.96 | 0.95 |
| 1 | 0.29 | 0.20 | 0.23 |
| XGBoost | 0 | 0.92 | 0.99 | 0.95 |
| 1 | 0.48 | 0.10 | 0.17 |







*Hình 4.1 Đồ thị biểu diễn đường cong ROC của các mô hình huấn luyện trên toàn bộ dữ liệu*

*Bảng 4.5 Kết quả tổng hợp chỉ số đánh giá mô hình*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Mô hình** | **Accuracy rate (%)** | **Recall rate (%)** | **Precision**  **Rate (%)** | **Gini** | **F1 Score** | **ROC**  **AUC** | **K-S**  **Statistic** |
| Logistic regression | 0.92 | 0.99 | 0.92 | 0.46 | 0.88 | 0.73 | 0.2 |
| Random forest | 0.92 | 1.0 | 0.92 | 0.38 | 0.88 | 0.69 | 0.17 |
| XGBoost | 0.92 | 0.99 | 0.92 | 0.5 | 0.88 | 0.75 | 0.23 |

Vì dữ liệu có sự mất cân bằng lớn khi tỷ lệ giá trị phân loại là 0 (khách hàng không có nguy cơ vỡ nợ) chiếm tỷ lệ lớn trong bộ dữ liệu gây khó khăn cho việc mô hình có thể dự đoán tốt cho giá trị phân loại là 1 (khách hàng có nguy cơ vỡ nợ) nên với các chỉ số đánh giá trên từng biến phân loại (*Bảng 4.4*) sẽ quan tâm nhiều hơn đến các chỉ số đánh giá trên giá trị phân loại là 1. Từ các chỉ số đánh giá mô hình dự đoán trên 2 biến phân loại có thể thấy mô hình XGBoost cho chất lượng phân loại tốt nhất trên các giá trị phân loại là 1 đồng nghĩa với việc mô hình này bao quát tốt được thông tin từ các thuộc tính của biến mục tiêu phân loại là 1 trong bộ dữ liệu. Các mô hình còn lại có có chỉ số đánh giá chất lượng dự đoán giá trị phân loại là 1 chưa tốt, các mô hình này có thể phải cần nhiều hơn dữ liệu có giá trị phân loại là 1.

Khi đánh giá chung các mô hình dựa trên tất cả các giá trị dự đoán trong *bảng 4.5* cũng cho thấy mô hình XGBoost có chất lượng các kết quả dự đoán là tốt nhất khi phần lớn các chỉ số đánh giá chất lượng của mô hình này cao hơn so với các mô hình còn lại.

Chỉ số ROC AUC cho mức cao nhất ở mô hình XGBoost là 0.75 Chỉ số này thấp hơn ở các mô hình Logistic regression, Random forest với giá trị thấp nhất ở mô hình Random forest (0.69).

Từ các chỉ số đánh giá chất lượng của các mô hình được huấn luyện trên toàn bộ dữ liệu, nhìn chung, XGBoost có kết quả tốt và đạt được kỳ vọng đề ra Các mô hình, Logistic regression, Random forest cho kết quả tốt trên một số chỉ số như accuracy rate, tuy nhiên các chỉ số còn lại khá thấp, có thể các mô hình đã khái quát tốt các thông tin về dữ liệu của các giá trị phân lớp là 0 nhưng chưa tốt trên các giá trị phân lớp là 1 điều này đến từ việc mất cân bằng của dữ liệu đầu vào.

### 4.2.2. Điểm tín dụng

Khi đã xây dựng được các mô hình và đánh giá sẽ chọn ra mô hình có kết quả tốt nhất để dự đoán, các giá trị này sẽ được chuyển đổi thành điểm tín dụng dựa trên phương pháp chuyển đổi điểm tín dụng (*2.6.5*). Trong phần này cũng đưa ra các dự đoán lớp phân loại của mô hình với ngưỡng phân loại là 0.6 được tính toán dựa trên phương pháp tìm ngưỡng phân loại (*2.6.4*) và các giá trị phân loại thực tế.

Kết quả của một số dữ liệu được dự đoán và chuyển đổi thành điểm tín dụng được trình bày dưới đây:

*Bảng 4.13 Giá trị dự đoán của mô hình XGBoost được chuyển đổi thành điểm tín dụng của 10 mẫu dữ liệu với ngưỡng phân loại 0.6*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Xác suất dự đoán** | **Nhãn**  **thực tế** | **Nhãn dự đoán** | **Điểm tín dụng** |
| 1 | 0.67 | 1 | 1 | 483 |
| 2 | 0.66 | 1 | 1 | 488 |
| 3 | 0.61 | 1 | 1 | 513 |
| 4 | 0.49 | 1 | 0 | 581 |
| 5 | 0.12 | 1 | 0 | 783 |
| 6 | 0.03 | 0 | 0 | 832 |
| 7 | 0.05 | 0 | 0 | 825 |
| 8 | 0.05 | 0 | 0 | 821 |
| 9 | 0.05 | 0 | 0 | 822 |
| 10 | 0.07 | 0 | 0 | 812 |

Trong 10 mẫu dữ liệu đã đưa ra có 5 mẫu mang nhãn phân lớp là 0 và 5 mẫu mang nhãn phân lớp là 1. Mô hình đã phân lớp đúng hầu hết các mẫu dữ liệu với ngưỡng phân loại tìm được dựa trên phương pháp tìm ngưỡng phân loại.

### 4.2.3. Thuộc tính quan trọng

Ngoài kết quả dự đoán, các mô hình sẽ đưa các thông tin về các thuộc tính quan trọng ảnh hưởng đến kết quả phân loại. Các thuộc tính dưới đây được tổng hợp từ mô hình tốt nhất là XGBoost.

*Bảng 4.7 Các thuộc tính quan trọng của mô hình*

|  |  |
| --- | --- |
| **Thuộc tính** | **Chú thích** |
| APP\_NAME\_EDUCATION\_TYPE\_Higher\_education | Khách hàng cấp giáo dục bậc cao |
| APP\_EXT\_SOURCE\_2 | Phân nhóm theo  ext\_source\_1 |
| APP\_EXT\_SOURCE\_3 | Phân nhóm theo  ext\_source\_2 |
| APP\_NAME\_INCOME\_TYPE\_Working | Khách hàng có nguồn thu nhập từ lao động |
| APP\_NAME\_INCOME\_TYPE\_Pensioner | Khách hàng có nguồn thu nhập từ hưu trí |
| APP\_CODE\_GENDER | Giới tính khách hàng |
| PREV\_INS\_NEW\_NUM\_PAID\_LATER\_SUM | Số lần trả muộn khoản trả góp |
| CCB\_CNT\_DRAWINGS\_ATM\_CURRENT\_MEAN | Trung bình số tiền rút từ  ATM |

|  |  |
| --- | --- |
| PREV\_NAME\_CONTRACT\_STATUS\_Refused\_MEAN | Trung bình số lượng hợp đồng xoay vòng |
| APP\_NAME\_EDUCATION\_TYPE\_Secondary | Khách hàng có cấp độ giáo dục cao nhất là giáo dục trung học |
| APP\_FLAG\_DOCUMENT\_3 | Khách hàng cung cấp tài liệu loại 3 |
| PREV\_INS\_AMT\_PAYMENT\_MIN\_SUM | Tổng số tiền khách hàng thanh toán trả góp mức tối thiểu |
| APP\_EXT\_SOURCE\_1 | Phân nhóm theo  ext\_source\_3 |
| APP\_FLAG\_OWN\_CAR | Khách hàng sở hữu ô tô |
| BB\_MONTHS\_BALANCE\_COUNT | Số tháng có số dư tín dụng |
| BB\_NEW\_CLOSED\_AMT\_ANNUITY\_MEAN | Số tiền trung bình hàng năm của các hợp đồng có trạng thái đã đóng |
| APP\_AMT\_REQ\_CREDIT\_BUREAU\_QRT | Số lượng thắc mắc gửi về phong tín dụng của khách hàng trước 3 tháng nộp  đơn |
| APP\_FLAG\_EMP\_PHONE | Khách hàng cung cấp số điện thoại cơ quan |
| APP\_REGION\_RATING\_CLIENT\_W\_CITY | Xếp hạng khu vực sinh sống tính đến thành phố |
| BB\_AMT\_CREDIT\_SUM\_OVERDUE\_SUM | Tổng số tiền khách hàng trả muộn |

## 4.3. Kết luận

Sau quá trình thực nghiệm các mô hình trên dữ liệu Home Credit Default Risk đã qua xử lý và trích xuất đặc trưng. Một số kết luận được rút ra như sau:

* Quá trình trích xuất đặc trưng từ dữ liệu gốc là quá trình rất quan trọng.

Đây là nền tảng chất lượng cho quá trình huấn luyện mô hình.

* Mô hình XGBoost có chất lượng tốt nhất trong các mô hình dự đoán.
* Do dữ liệu mất cân bằng nên mô hình Logistic regression và Random forest không đạt kết quả cao.

# CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## 5.1. Kết quả

### 5.1.1. Kết quả đạt được

#### 5.1.1.1. Kết quả bài toán

Sau quá trình nghiên cứu và thực nghiệm đã xây dựng thành công các mô hình chấm điểm tín dụng dựa trên các thuật toán nổi bật trong học máy và đưa ra được mô hình tốt nhất với chỉ số đánh giá độ chính xác xấp xỉ 80% và có thứ hạng cao trên bảng xếp hạng hiệu suất mô hình của cuộc thi được tổ chức. Mô hình chấm điểm tín dụng đưa ra có kết quả khá tốt và cho thấy mức độ khả quan cao về việc ứng dụng mô hình vào bối cảnh thực tế mà đề tài đã đưa ra từ phần tổng quan nghiên cứu. Ngoài ra, mô hình còn cho thấy tiềm năng cao khi ứng dụng được với các mô hình chấm điểm tín dụng truyền thống, điều này có ý nghĩa thực tiễn rất lớn giúp cho các ngân hàng, các định chế tài chính, các tổ chức tín dụng có thể đánh giá triệt để được các rủi ro tín dụng có liên quan đến hoạt động cho vay.

#### 5.1.1.2. Kiến thức và kỹ năng

Trong suốt quá trình nghiên cứu và thực hiện đề tài này, bản thân đã đạt được một số kết quả:

* Hiểu được tầm quan trọng của bài toán chấm điểm tín dụng và các kiến thức chuyên ngành tài chính tín dụng liên quan.
* Nắm được các lý thuyết về các xây dựng và triển khai spark và hadoop.
* Nắm được các lý thuyết về các phương pháp, kỹ thuật của các thuật toán Logistic regression, Random forest, XGBoost.
* Thực nghiệm xây dựng và so sánh được các mô hình dữ liệu từ các thuật toán trên.
* Trong quá trình thực hiện luận văn đã giúp bản thân nâng cao được khả năng đọc hiểu và tham khảo các bài nghiên cứu khoa học từ đó giúp cung cấp được những nền tảng kiến thức trong quá trình thực hiện đề tài.
* Bản thân cũng nâng cao được khả năng trình bày, soạn thảo và báo cáo các nội dung nghiên cứu và thực nghiệm.

## 5.2. Hướng phát triển

Việc áp dụng các thuật toán học máy vào mô hình chấm điểm tín dụng cho thấy kết quả khả thi và mang tính ứng dụng cao. Trong tương lai, ngoài những kết quả đã nghiên cứu và thực nghiệm sẽ định hướng phát triển đề tài với các mục tiêu:

* Nghiên cứu thêm về các phương pháp khác có liên quan đến xây dựng mô hình chấm điểm tín dụng và các thuật toán học máy. Bên cạnh đó tiếp tục cập nhật các phương pháp thuật toán mới nhằm đưa ra được mô hình chất lượng hơn.
* Thực nghiệm các thuật toán với một số bộ dữ liệu khác có liên quan đến cho vay tín dụng như *Lending Club Loan Data* [73], *default of credit card clients* [74], … từ đó nghiên cứu về mô hình chấm điểm tín dụng với quy mô rộng hơn và khái quát hơn về nhiều mẫu dữ liệu khác nhau.
* Triển khai mô hình đã huấn luyện lên nền tảng web để phù hợp hơn với

định hướng phát triển mô hình cho người dùng sử dụng.

# TÀI LIỆU THAM KHẢO

1. "Tạp chí tài chính," 21 11 2021. [Online]. Available:

https://tapchitaichinh.vn/giai-phap-tin-dung-va-xu-ly-no-xau-trong-dieukien-nen-kinh-te-bi-tac-dong-boi-dai-dich-covid-19.html.

1. G. T. T. Huyền, "Một số kĩ thuật học máy cho chấm điểm tín dụng," *Tạp chí Khoa học & Đào tạo Ngân hàng,* no. 227, pp. 34-40, 2021.
2. M. A. Qureshi, "Credit Scoring and Its Applications in Banking: A Comprehensive Review," *Journal of Risk and Financial Management,* 2020.
3. https://www.investopedia.com/terms/f/ficoscore.asp, "Investopedia," Fair Isaac Corporation, 18 February 2023. [Online]. Available:

https://www.investopedia.com/terms/f/ficoscore.asp.

1. H. Sargeant, "Algorithmic decision-making in financial services: economic and normative outcomes in consumer credit," *AI and Ethics,* 2022.
2. "The Council on Foreign Relations," 19 February 2015. [Online]. Available: https://www.cfr.org/backgrounder/credit-rating-controversy.
3. S. Khemakhem and Y. Boujelbene, "Predicting credit risk on the basis of financial and non-financial variables and data mining," *Review of Accounting and Finance,* 2018.
4. "Home Credit Default Risk," Group Home Credit, [Online]. Available: https://www.kaggle.com/competitions/home-credit-default-risk/data.
5. T. Tamplin, "Finance Strategists," Finance Strategists, March 2023. [Online].

Available:

https://www.financestrategists.com/banking/creditscore/?gclid=Cj0KCQjw0tKiBhC6ARIsAAOXutmxWG5jpfc1eJ3dfaBte6P eMwYtsbilgOatm1KD\_EPoUc91\_yytbq4aAsvgEALw\_wcB.

1. P. G. M. K. Brevoort Kenneth, "Data Point: Credit Invisibles," *Consumer Financial Protection Bureau,* 2015.
2. F. R. B. o. Philadelphia, "The Long-Term Effects of Credit Scores on Credit Market Outcomes," 2015.
3. S. J. Hand David, Statistics in Finance (Arnold Applications of Statistics Series), London: Wiley, 1998.
4. "FICO," Fair Isaac Corporation, [Online]. Available:

https://www.fico.com/en/products/fico-score.

1. A. Hayes, "investopedia," February 2023. [Online]. Available:

https://www.investopedia.com/terms/f/ficoscore.asp.

1. B. Fav, "What is a Credit Score & How is it Calculated?," America's Debt Help, [Online]. Available:

https://www.debt.org/credit/report/scoring-models/.

1. "The VantageScore Model - Empowering lenders to make confident and inclusive lending decisions.," VantageScore, 2023. [Online]. Available: https://www.vantagescore.com/lenders/why-vantagescore/our-models/.
2. G. Z. J. L. Y Zhu, "The impact of alternative data sources on credit scoring models," *Journal of Business Research,* pp. 205-215, 2019.
3. Y. L. X. L. H Lin, "Incorporating social media data into credit scoring models," *Decision Support Systems,* pp. 78-88, 2019.
4. B. Bhattacharya, "Credit risk assessment: The impact of macroeconomic factors," *Journal of Risk Research,* pp. 1098-1119, 2019.
5. Apache Spark Documentation. https://spark.apache.org/
6. Zaharia, M. et al. (2016). Spark: The Definitive Guide. O'Reilly Media.
7. Karau, H. et al. (2017). Learning Spark: Lightning-Fast Big Data Analysis. O'Reilly Media.
8. White, Tom. *Hadoop: The Definitive Guide*. O'Reilly Media, 2015.
9. Apache Software Foundation. *"What is Apache Hadoop?"* <https://hadoop.apache.org>
10. Vervest, Peter, et al. *"MapReduce and Hadoop for Large Scale Data Processing."* *Journal of Big Data*, vol. 10, no. 2, 2020, pp. 56-69.
11. IBM. *"Overview of Hadoop and Its Ecosystem."* <https://www.ibm.com>
12. Cloudera. *"Understanding Hadoop’s Distributed File System (HDFS)."* <https://www.cloudera.com>
13. "IBM," [Online]. Available: https://www.ibm.com/topics/machine-learning.
14. R. W. Picard, E. Vyzas and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence,* vol. 23, pp. 1175-1191, 2001.
15. W. D. W. D. Luo Cuicui, "A deep learning approach for credit scoring using credit default swaps," *Engineering Applications of Artificial Intelligence,* vol. 65, pp. 465-470, 2017.
16. H. Z. W. S. J Jiang, "Credit risk assessment based on neural networks," *International Journal of Simulation: Systems, Science and Technology,* vol. 19(14), pp. 1-8, 2018.
17. A. T. Bouzouita, "A machine learning approach to credit scoring: Evidence from SMEs," *Journal of Business Research,* vol. 117, pp. 807-816, 2020.
18. D. Zhang, X. Zhou, S. C. Leung, Zheng and Jiemin, "Vertical bagging decision trees model for credit scoring," *Expert Systems with Applications,* vol. 37, pp. 7838-7843, 2010.
19. D. Lachos-Perez, S. Dussán-Sarria and L. Giraldo-Gómez, "Bayesian variable selection and classification model for credit scoring," *Expert Systems with Applications,* 2021.
20. LaValley and M. P, "Logistic regression," *Circulation,* vol. 117, pp. 23952399, 2008.
21. C. Gavin, T. Nicola and G. Mark, "Sparse multinomial logistic regression via bayesian l1 regularisation," *Advances in neural information processing systems,* vol. 19, 2006.
22. L. Rokach, "Ensemble methods for classifiers," *Data mining and knowledge discovery handbook,* pp. 957-980, 2005.
23. S. Omer and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 8, p. 1249, 2018.
24. F. Khaled , M. M. Gaber and E. Elyan, "Random forests: from early developments to recent advancements," *Systems Science & Control Engineering: An Open Access Journal,* vol. 2, pp. 602-609, 2014.
25. D. G. Thomas, "Ensemble learning," *The handbook of brain theory and neural networks,* vol. 2, pp. 110-125, 2002.
26. R. Polikar, "Ensemble learning," *Ensemble machine learning: Methods and applications,* pp. 1-34, 2012.
27. P. Sollich and A. Krogh, "Learning with ensembles: How overfitting can be useful," *Advances in neural information processing systems,* vol. 8, 1995.
28. N. Hu and R. B. Dannenberg, "Bootstrap learning for accurate onset detection," *Machine Learning,* vol. 65, pp. 457-471, 2006.
29. B. Boehmke and B. M. Greenwell, Hands-on machine learning with R, CRC press, 2019.
30. K. M. Mendez, S. N. Reinke and D. I. Broadhurst, "A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification," *Metabolomics,* vol. 15, pp. 1-15, 2019.
31. T.-H. Lee, A. Ullah and R. Wang, "Bootstrap aggregating and random forest," *Macroeconomic forecasting in the era of big data: Theory and practice,* pp. 389-429, 2020.
32. T. K. Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence,* vol. 20, pp. 832-844, 1998.
33. "github," 6 2022. [Online]. Available: https://github.com/dmlc/xgboost.
34. G. Kalipe, V. Gautham and R. K. Behera, "Predicting malarial outbreak using machine learning and deep learning approach: a review and analysis," *2018 International Conference on Information Technology (ICIT),* pp. 33-38, 2018.
35. M. Piryonesi and T. El-Diraby, "Data Analytics in Asset Management: CostEffective Prediction of the Pavement Condition Index," *Journal of Infrastructure Systems,* vol. 26, 2020.
36. T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani and J. Friedman, "Boosting and additive trees," *The elements of statistical learning: data mining, inference, and prediction,* pp. 337-387, 2009.
37. P. Madeh and T. El-Diraby, "Using machine learning to examine impact of type of performance indicator on flexible pavement deterioration modeling," *Journal of Infrastructure Systems,* vol. 27, 2021.
38. J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu and Y. Xiang, "Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China," *Energy conversion and management,* vol. 164, pp. 102-111, 2018.
39. R. C. Turner, A. Fuggetta, L. Lavazza and A. L. Wolf, "A conceptual basis for feature engineering," *Journal of Systems and Software,* vol. 49, pp. 3-15, 1999.
40. Z. Guoping, "A Necessary Condition for a Good Binning Algorithm in Credit Scoring," *Applied Mathematical Sciences,* vol. 8, pp. 3229-3242, 2014.
41. W. Wang, C. Lesner, A. Ran, M. Rukonic, J. Xue and E. Shiu, "Using small business banking data for explainable credit risk scoring," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 13396-13401.
42. H. A. Abdou, "Genetic programming for credit scoring: The case of Egyptian public sector banks," *Expert systems with applications,* vol. 36, pp. 1140211417, 2009.
43. S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv,* 2016.
44. J. Daniel and M. H. James, "Logistic Regression," in *Speech and Language Processing*, California, Stanford University Press, 2023.
45. J. Xu, Y. Zhang and D. Miao, "Three-way confusion matrix for classification:

A measure driven view," *Information sciences,* vol. 507, pp. 772-794, 2020.

1. Y. Li, T. Bellotti and N. Adams, "Issues using logistic regression with class imbalance, with a case study from credit risk modelling," *Foundations of Data Science,* vol. 1, pp. 389-417, 2019.
2. A. Luque, A. Carrasco, A. Mart and A. de Las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition,* vol. 91, pp. 216-231, 2019.
3. J. Brownlee, "Machine Learning Mastery," August 2020. [Online]. Available: https://machinelearningmastery.com/confusion-matrix-machine-learning/.
4. T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters,* vol. 27, pp. 861-874, 2006.
5. M. Vuk and T. Curk, "ROC curve, lift chart and calibration plot," *Advances in methodology and Statistics,* vol. 3\, pp. 89-108, 2006.
6. S. Leteurtre, A. Martinot, A. Duhamel, F. Proulx, B. Grandbastien, J. Cotting, R. Gottesman, A. Joffe, J. Pfenninger and P. Hubert, "Validation of the paediatric logistic organ dysfunction (PELOD) score: prospective, observational, multicentre study," *The Lancet,* vol. 362, pp. 192-197, 2003.
7. M. J. Miller, "Credit Reporting Systems and the International Economy," The MIT Press, Cambridge, 2003.
8. G. G. Grinstein and M. O. Ward, "Introduction to data visualization," *Information visualization in data mining and knowledge discovery,* vol. 1, pp. 21-45, 2002.
9. V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering,* vol. 2, pp. 42-47, 2012.
10. D. Marghescu, "Multidimensional data visualization techniques for financial performance data: A review," *Turku Centre for Computer Science,* 2007.
11. C. Jiang, Z. Wang, R. Wang and Y. Ding, "Loan default prediction by combining soft information extracted from descriptive text in online peer-topeer lending," *Annals of Operations Research,* vol. 266, pp. 511-529, 2018.
12. M. Zhang, W. Li, Q. Du, L. Gao and B. Zhang, "Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN," *IEEE transactions on cybernetics,* vol. 50, pp. 100-111, 2018.
13. "Kaggle," [Online]. Available: https://www.kaggle.com/docs/notebooks.
14. "Kaggle," [Online]. Available:

https://www.kaggle.com/datasets/wordsforthewise/lending-club.

1. "UCI Machine Learning Repository," [Online]. Available:
2. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

# NHẬT KÝ LÀM VIỆC

|  |  |  |  |
| --- | --- | --- | --- |
| **Tuần** | **Từ ngày** | **Đến ngày** | **Nội dung** |
| 1 | 03/09/2024 | 09/09/2024 | Đăng ký đề tài nghiên cứu.  Làm các hồ sơ thủ tục liên quan.  Họp bàn đề tài với giáo viên hướng dẫn. |
| 2 | 10/09/2024 | 16/09/2024 | Tìm hiểu về đề tài. Đưa ra các nội dung liên quan đến đề tài, hướng nghiên cứu. |
| 3 | 17/09/2024 | 23/09/2024 | Nghiên cứu các kiến thức liên quan đến hoạt động tín dụng cho vay.  Học thêm các kỹ thuật liên quan đến tài chính ngân hàng nhằm hiểu thêm về đề tài. |
| 4 | 24/09/2024 | 30/10/2024 | Tiếp tục tìm hiểu đề tài thông qua các nghiên cứu có liên quan. |
| 5 | 01/10/2024 | 07/10/2024 | Tìm hiểu nghiên cứu sâu vào các thuật toán học máy và kỹ thuật liên quan để tiếp cận đến ứng dụng của thuật toán vào chấm điểm tín dụng. |
| 6 | 08/10/2024 | 13/10/2024 | Tìm kiếm dữ liệu phù hợp, đọc các tài liệu, các cuộc thảo luận liên quan nhằm tận dụng triệt để các thông tin dữ liệu mang lại, các kiến trúc mô hình để xây dựng xử lý dữ liệu lớn bằng spark. |
| 7 | 14/10/2024 | 20/10/2024 | Tiến hành các bước xử lý dữ liệu, phân tích và khám phá dữ liệu. |
| 8 | 21/10/2024 | 27/10/2024 | Xây dựng mô hình Logistic regression và Random forest trên bộ dữ liệu đã xử lý.  Đánh giá các mô hình. |

|  |  |  |  |
| --- | --- | --- | --- |
| 9 | 28/10/2024 | 04/11/2024 | Kết quả huấn luyện mô hình Logistic regression và Random forest không tốt và cần tìm hướng xử lý mới.  Trình bày vấn đề với giáo viên hướng dẫn, họp bàn tìm hướng giải quyết vấn đề |
| 10 | 05/11/2024 | 11/11/2024 | Sau khi trình bày vấn đề với giáo viên hướng dẫn được thông tin về việc cần có quá trình trích xuất đặc trưng từ dữ liệu.  Tiến hành quá trình trích xuất đặc trưng từ các kiến thức liên quan đến cho vay tín dụng xây dựng được các thuộc tính mới sát hơn với dữ liệu tín dụng thực tế. |
| 11 | 12/11/2024 | 18/11/2024 | Xây dựng lại mô hình Logistic regression và Random forest. Kết quả đánh giá cho thấy sự khả quan hơn. |
| 12 | 18/11/2024 | 24/11/2024 | Tiến hành xây dựng Hadoop,Spark,Airflow |
| 14 | 24/11/2024 | 30/11/2024 | Từ thông tin của GVHD, tìm hiểu và xây dựng mô XGBoost. Song song với đó tiếp tục tiến hành xây dụng spark và airflow để trainning với dữ liệu lớn |
| 15 | 31/11/2024 | 6/12/2024 | Các mô hình đã có kết quả nhưng chưa đạt được kỳ vọng đặt ra nên tiếp tục tìm hiểu thêm các thông tin liên quan có thể cải tiến chất lượng. Nhưng phần thiết bị cá nhân không đáp ứng đủ để xây dụng spark, hadoop và airflow liên hệ với GVHD để tìm hướng giải quyết |
| 16 | 7/12/2024 | 11/12/2024 | Xây dựng lại spark, hadoop và airflow trực tiếp trên thiết bị phần cứng mà GVHD đã cung cấp |
| 18 | 11/12/2024 | 15/12/2024 | Đánh giá tổng quan quá trình nghiên cứu và thực nghiệm.  Tổng hợp các thông tin có trong quá trình nghiên cứu và thực nghiệm.  Hoàn thiện báo cáo và bài thuyết trình. |