ĐẠI HỌC CÔNG NGHIỆP TP.HCM
KHOA CÔNG NGHỆ THÔNG TIN

KHÓA LUẬN TỐT NGHIỆP

MÔ HÌNH CHẤM ĐIỂM TÍN DỤNG SỬ DỤNG CÁC THUẬT TOÁN HỌC MÁY

CHUYÊN NGÀNH: KHOA HỌC DỮ LIỆU

Sinh viên thực hiện: Hồ Quang Huy
Giảng viên hướng dẫn: TS. Nguyễn Chí Kiên

01

TỔNG QUAN ĐỀ TÀI

Điểm tín dụng...

- Thang điểm số **từ 300 đến 850**.
- Hand & Jacka: "**Chấm điểm tín dụng** là quá trình (của các tổ chức tài chính) **lập mô hình về mức độ tin cậy"**.
- Một số mô hình chấm điểm tín dụng phổ biến:
  - Mô hình chấm điểm tín dụng **FICO** (1989)
  - Mô hình chấm điểm **VantageScore** (2006)

[David J. Hand, 1998, Statistics in Finance]

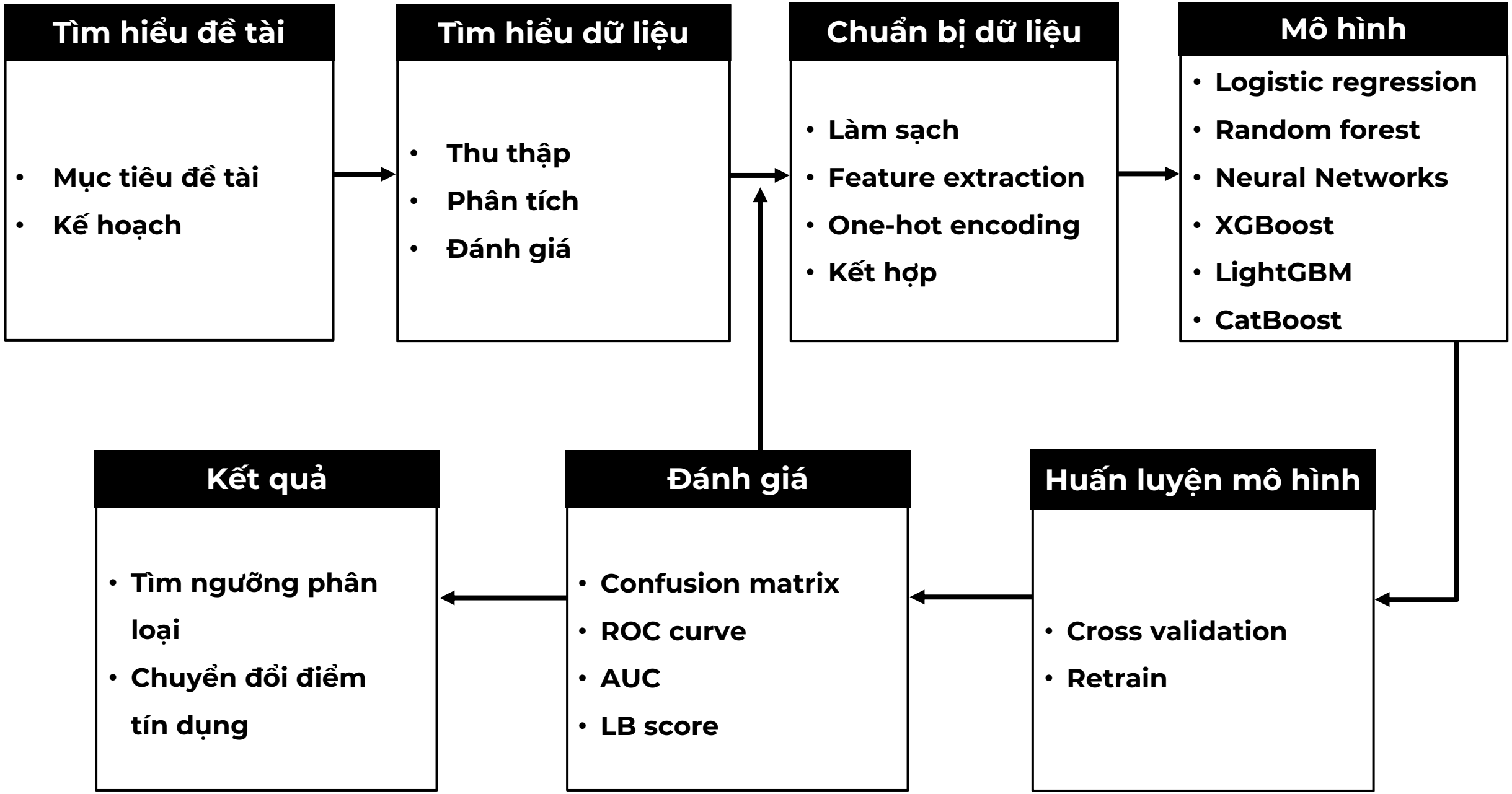
Đặt vấn đề

- **Mô hình chấm điểm tín dụng truyền thống** thường dựa vào các thuộc tính **cố định và hạn chế** - không tính đến các yếu tố **phi tài chính** (Khủng hoảng tài chính 2008).
- Mô hình sử dụng **thuật toán học máy** có thể phân tích **khối lượng dữ liệu lớn**, bao gồm cả những yếu tố **phi tài chính**.

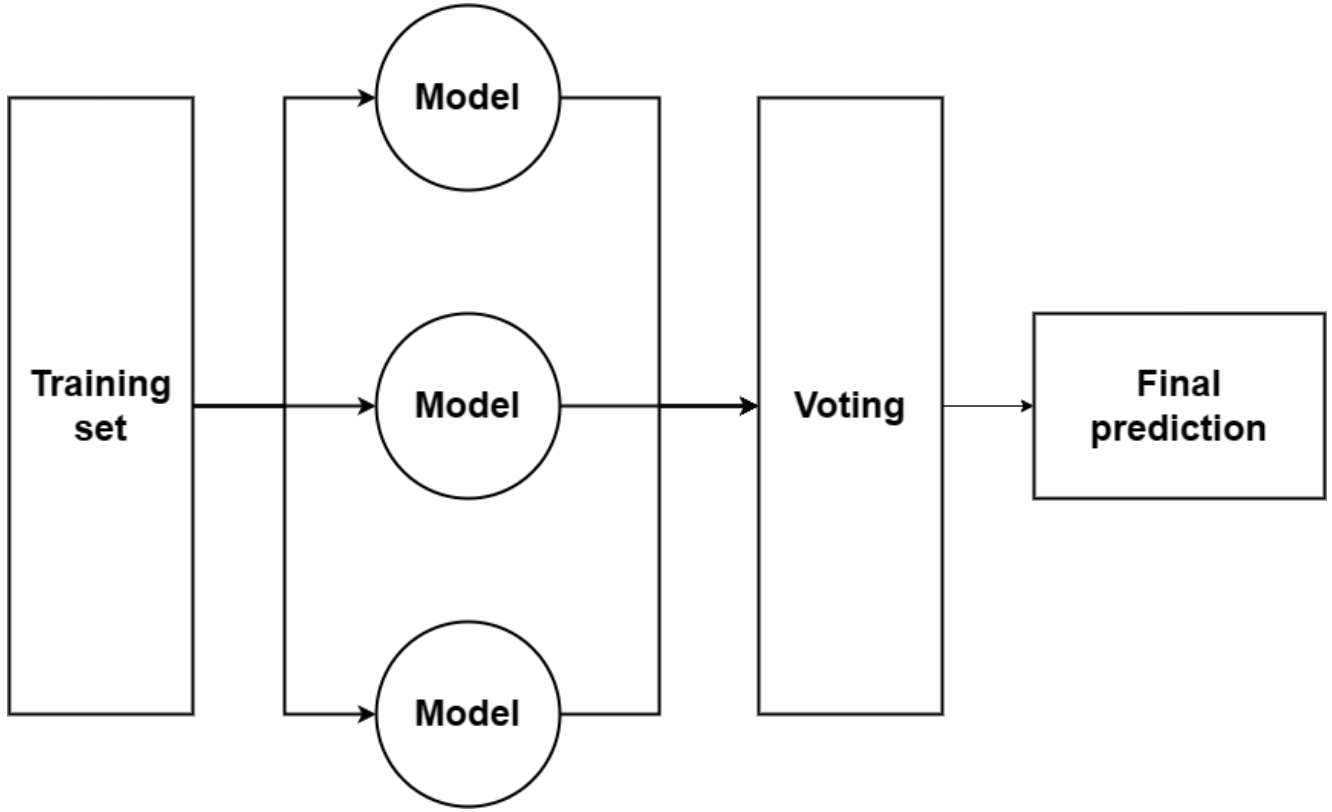
[Khemakhem et al., 2018, Predicting credit risk on the basis of financial and non-financial variables and data mining]

02

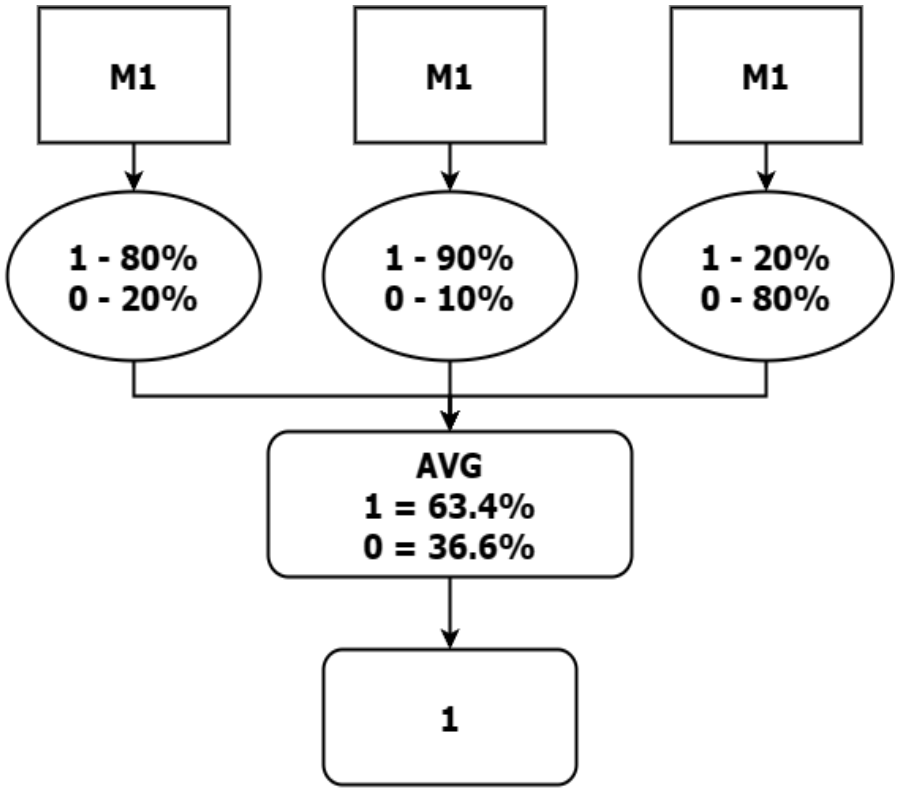
PHƯƠNG PHÁP


Tìm hiểu đề tài
Mục tiêu đề tài
Kế hoạch
Tìm hiểu dữ liệu
Thu thập
Phân tích
Đánh giá
Chuẩn bị dữ liệu
Làm sạch
Feature extraction
One-hot encoding
Kết hợp
Mô hình
Logistic regression
Random forest
Neural Networks
XGBoost
LightGBM
CatBoost
Huấn luyện mô hình
Cross validation
Retrain
Đánh giá
Confusion matrix
ROC curve
AUC
LB score
Kết quả
Tìm ngưỡng phân loại
Chuyển đổi điểm tín dụng

Học đồng bộ (Ensemble learning)



Quá trình xây dựng mô hình dựa trên phương pháp ***Voting***



Phương thức ***soft-voting***

03

THỰC NGHIỆM VÀ KẾT QUẢ

Home Credit Default Risk

- Tổng hợp và công khai bởi **Home Credit Group.**
- Thông tin liên quan đến **dữ liệu tín dụng, tài chính, phi tài chính**.
- Biến mục tiêu phân loại:
  - (**0**): Không có khả năng vỡ nợ (99%)
  - (**1**): Có khả năng vỡ nợ (1%)

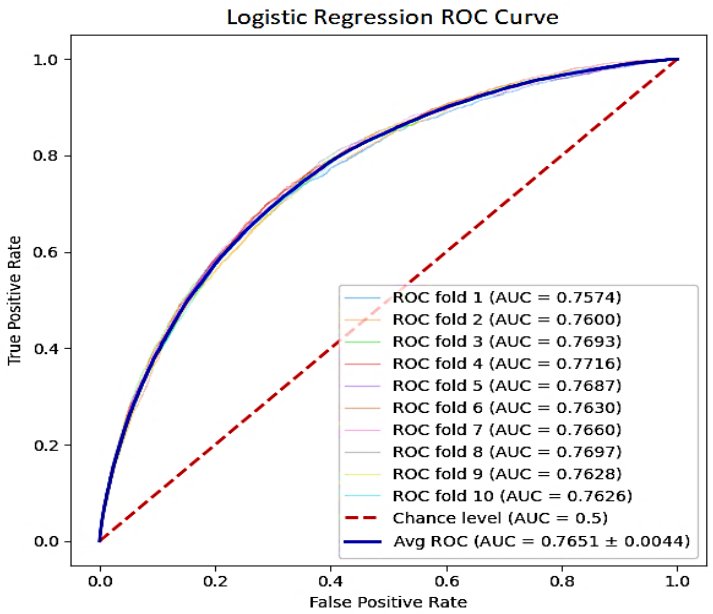
Chuẩn bị thực nghiệm

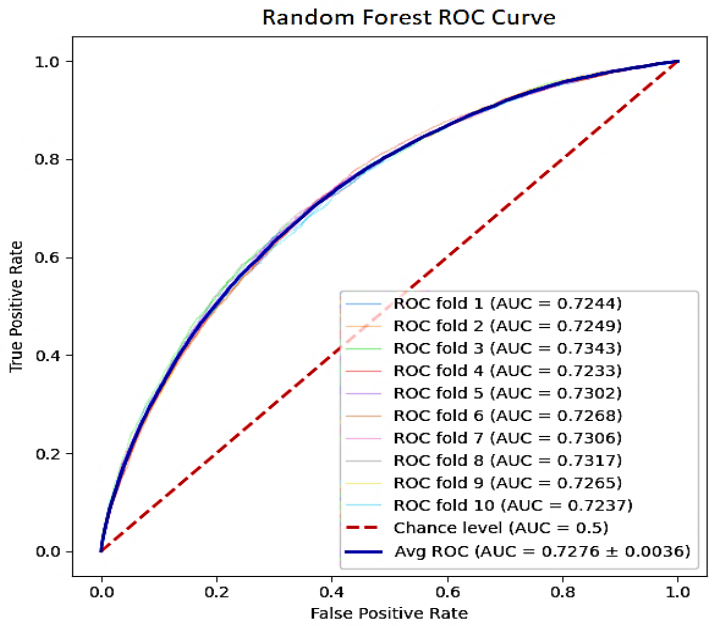
- Dữ liệu được kết hợp từ quá trình chuẩn bị dữ liệu:
  - **Train set**: 356251 mẫu và 781 thuộc tính.
  - **Test set**: 48744 mẫu và 781 thuộc tính.
- Cấu hình phần cứng sử dụng:
  - CPU: Intel Xeon 2-core
  - RAM: 13 GB
  - ROM: 73 GB
  - GPU: P100 - 16GB VRAM

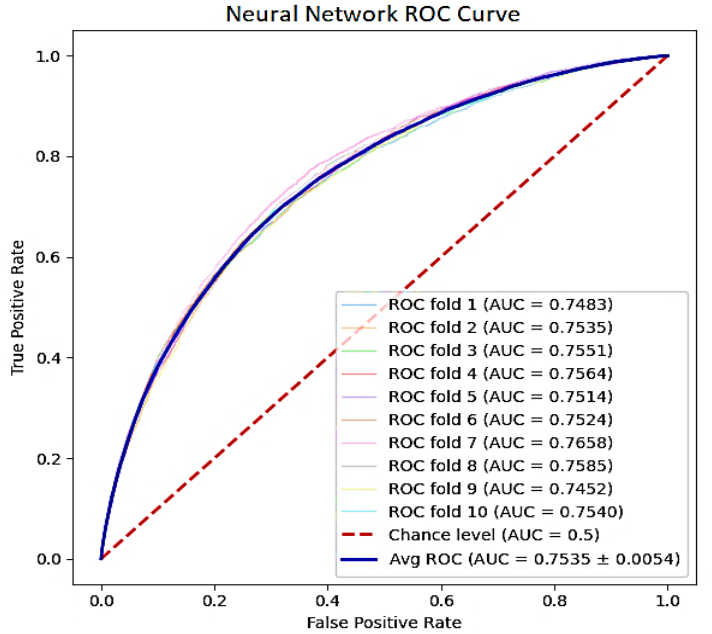
Thời gian huấn luyện

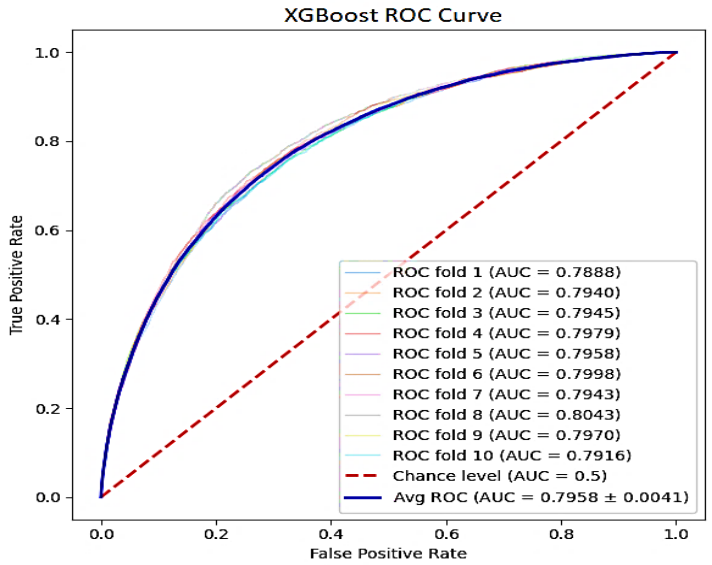
Mô hình	Thời gian huấn luyện
Logistic regression	2 tiếng 37 phút
Random forest	3 tiếng 54 phút
Neural network	4 tiếng 19 phút
XGBoost	6 tiếng 22 phút
LightGBM	3 tiếng 48 phút
CatBoost	5 tiếng 35 phút

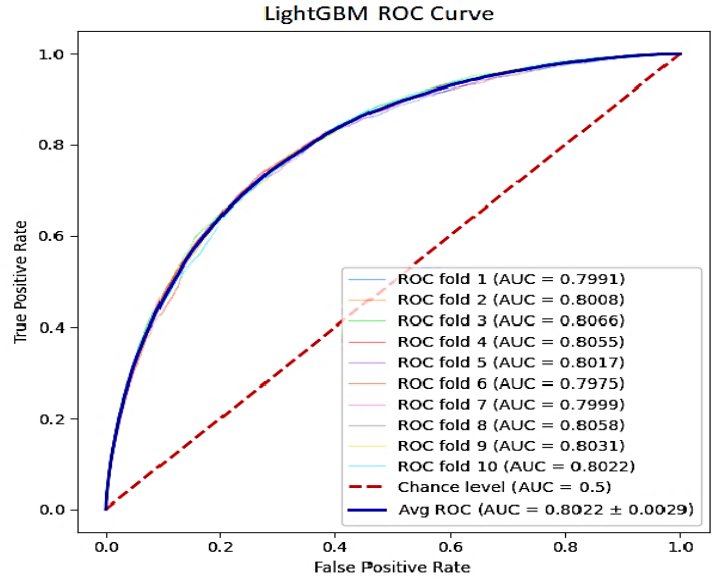
Huấn luyện cross-validation

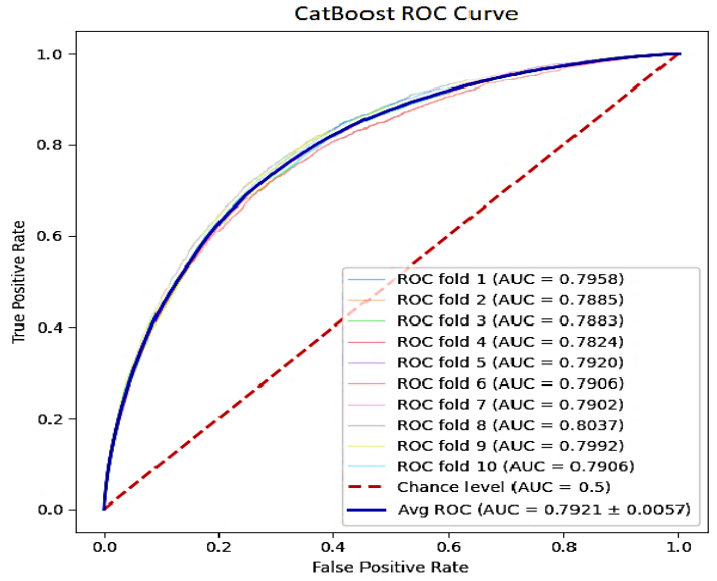












Kết quả huấn luyện cross-validation

Mô hình	$\mathbf{mean_{AUC}}$	$\mathbf{std_{AUC}}$
Logistic regression	0.76	0.0044
Random forest	0.72	0.0036
Neural network	0.75	0.0054
XGBoost	0.79	0.0041
LightGBM	0.80	0.0029
CatBoost	0.79	0.0057

Huấn luyện trên toàn bộ dữ liệu

Chỉ số đánh giá trên biến phân loại

Mô hình	Biến phân loại	Precision	Recall	F1-score
Logistic regression	0	0.95	0.82	0.88
	1	0.21	0.35	0.31
Random forest	0	0.93	0.96	0.94
	1	0.30	0.20	0.24
Neural network	0	0.92	1.00	0.96
	1	0.28	0.06	0.08
XGBoost	0	0.94	0.99	0.96
	1	0.45	0.09	0.16
LightGBM	0	0.93	1.00	0.96
	1	0.54	0.02	0.03
CatBoost	0	0.94	0.97	0.96
	1	0.57	0.32	0.27

Huấn luyện trên toàn bộ dữ liệu

Kết quả tổng hợp chỉ số đánh giá mô hình

Mô hình	Accuracy rate (%)	Recall rate (%)	Precision Rate (%)	Specificity Rate (%)	F1 Score	ROC AUC
Logistic regression	79.4673	22.0183	21.3050	81.6223	0.3078	0.7662
Random forest	89.5841	20.0632	30.1126	95.8220	0.2408	0.7323
Neural network	91.9287	13.795	27.9842	92.3761	0.2015	0.7528
XGBoost	92.8227	9.3738	45.1923	99.1395	0.1553	0.7959
LightGBM	92.983	33.6278	54.217	99.885	0.3475	0.7969
CatBoost	91.8122	55.4502	36.4597	97.0953	0.2746	0.7951

LB Score

Mô hình	Private score	Public score
Logistic regression	0.75805	0.76044
Random forest	0.71982	0.71719
Neural network	0.75303	0.75098
XGBoost	0.77623	0.77708
LightGBM	0.76562	0.77291
CatBoost	0.79095	0.79598

Kết quả huấn luyện mô hình học đồng bộ

Chỉ số đánh giá trên biến phân loại

Biến phân loại	Precision	Recall	F1-score
0	0.96	0.94	0.93
1	0.59	0.33	0.29

Tổng hợp chỉ số đánh giá

Accuracy rate (%)	Recall rate (%)	Precision Rate (%)	Specificity Rate (%)	F1 Score	ROC AUC	Private score	Public score	Training time
93.72	59.16	28.07	94.15	0.30	0.8293	0.7991	0.8095	8h13min

Điểm tín dụng

Chuyển đổi điểm tín dụng (ngưỡng phân loại 0.6)

STT	ID	Xác suất dự đoán	Nhãn thực tế	Nhãn dự đoán	Điểm tín dụng
1	439073	0.6089	1	1	515
2	415124	0.5912	1	0	524
3	347952	0.7674	1	1	427
4	158208	0.6917	1	1	469
5	453230	0.5842	1	0	528
6	183658	0.5405	0	0	552
7	144091	0.5102	0	0	569
8	139861	0.4907	0	0	580
9	199727	0.2175	0	0	730
10	236737	0.0251	0	0	836

05

KẾT LUẬN

Kết luận

- Quá trình trích xuất đặc trưng **rất quan trọng**.
- Mô hình **CatBoost có chất lượng tốt nhất**.
- LightGBM và XGBoost cũng có kết quả tốt.
- Mô hình LightGBM ⟶ **tối ưu thời gian huấn luyện** nhưng vẫn mang lại **hiệu suất mô hình đủ tốt**.
- Logistic regression và Random forest không đạt kết quả cao.
- Deep learning **không phù hợp** với loại dữ liệu này.
- Ensemble learning **cải thiện chất lượng** dự đoán của mô hình.

Hướng phát triển

- Nghiên cứu thêm về các **phương pháp**.

  (LSTM, Stacking-Meta Model, ...)

- Thực nghiệm với một số **bộ dữ liệu** khác.

  (*Lending Club Loan Data*, *default of credit card clients*, ...)

- Triển khai mô hình đã huấn luyện lên **nền tảng web**.

Thank you!