Trường Đại học Khoa học Tự nhiên, ĐHQGHN
Khoa Toán – Cơ – Tin học

Dự đoán khả năng chấp thuận cấp thẻ tín dụng dựa trên hồ sơ đăng kí

Đào Thị Ngọc Bích - 23001501
Nguyễn Hữu An - 23001493
Nguyễn Thị Quỳnh Anh - 23001496

Nội dung chính

1. Tổng quan dữ liệu
2. Tiền xử lý dữ liệu
3. Giảm chiều dữ liệu
4. Phân cụm dữ liệu
5. Xây dựng và đánh giá mô hình
6. Mở rộng: Bài toán Hồi quy

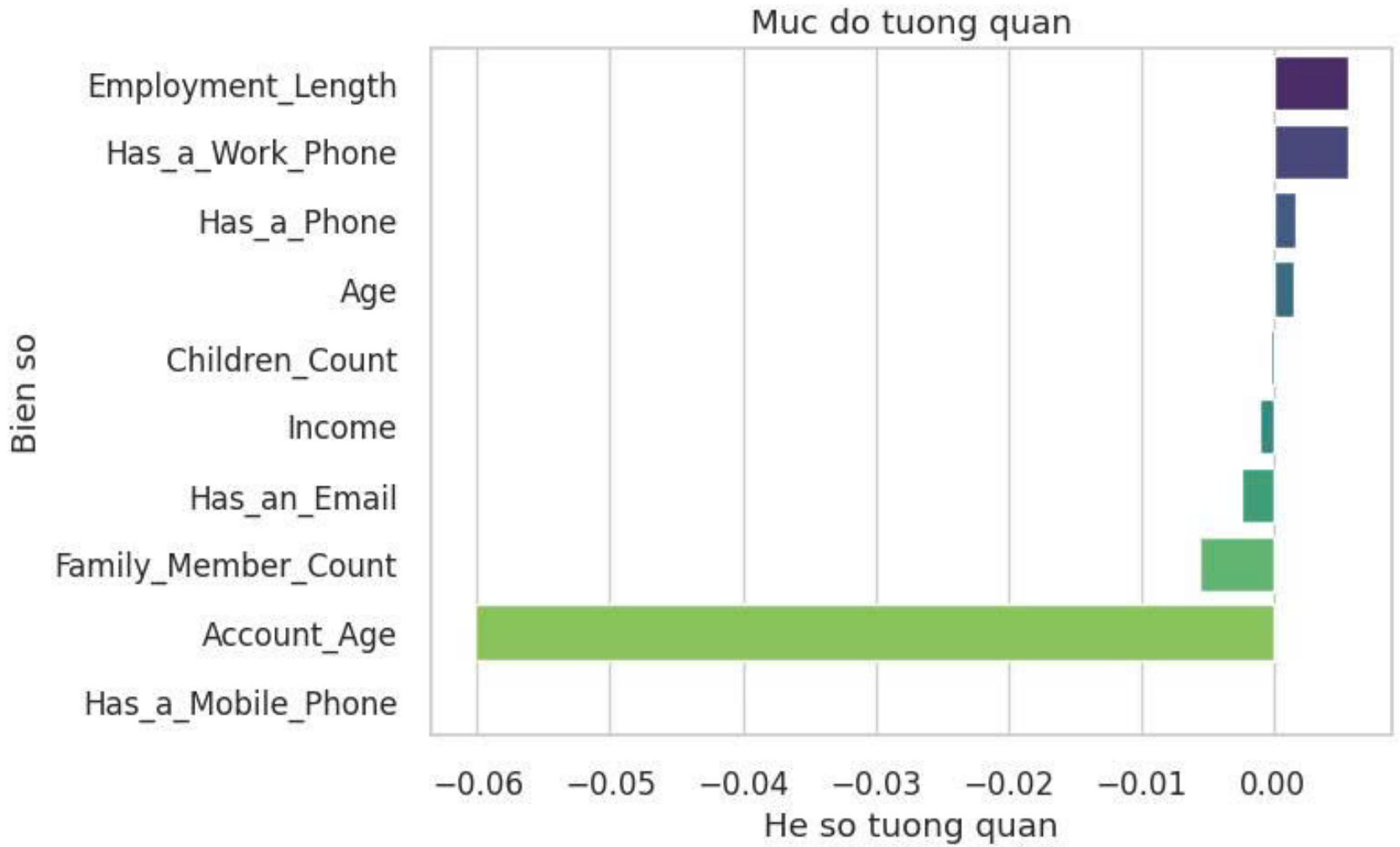
Nội dung

Tổng quan dữ liệu

Thông tin dataset

- Bộ dữ liệu gồm 36.457 bản ghi với 20 thuộc tính ban đầu.
- Biến mục tiêu: `Is_high_risk` (0: An toàn, 1: Rủi ro).
- Mục tiêu: Phân loại hồ sơ khách hàng có rủi ro nợ xấu hay không.

Phân tích tương quan



Hình 2.3: Tương quan với biến đầu ra

Nội dung

Loại bỏ thuộc tính không phù hợp

Rò rỉ dữ liệu và dữ liệu thiếu

- **Account_Age**: Loại bỏ do rò rỉ dữ liệu (Data Leakage), vì biến mục tiêu được xây dựng dựa trên lịch sử này.
- **Job_Title**: Loại bỏ do thiếu dữ liệu nghiêm trọng ($\approx 31\%$).

Thiếu phương sai và sức mạnh dự đoán

- **Has_a_Mobile_Phone**: Loại bỏ vì 100% khách hàng đều có (Variance $= 0$).
- **Children_Count**: Loại bỏ do sức mạnh dự đoán thấp, không có sự khác biệt rõ rệt về tỷ lệ rủi ro.

Chuyển đổi dữ liệu và Thống kê

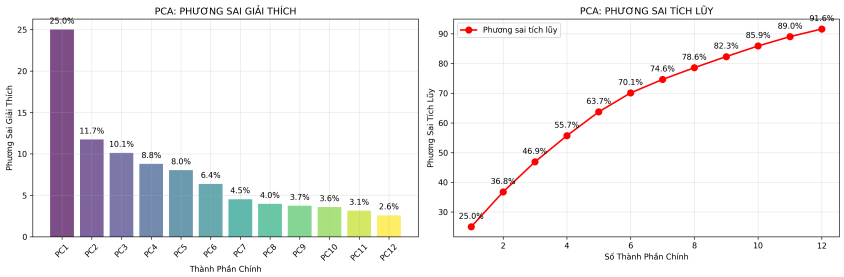
- **Outliner Handling**: Family_Member_Count, Income, Employment_Length
- **One-Hot Encoding**: Gender, Dwelling, Has_a_Car, Has_a_Property, ...
- **Ordinal Encoding**: Education Level
- **MaxMin Scaling**: Income, Age, ...

Nội dung

Phân tích thành phần chính (PCA)

Kết quả PCA

- Cần khoảng 12 thành phần chính để giải thích >90% phương sai.
- PC1: 25.02%, PC2: 11.74% phương sai.
- Hạn chế: Các điểm dữ liệu rủi ro và an toàn vẫn trộn lẫn, không tối ưu cho phân loại.

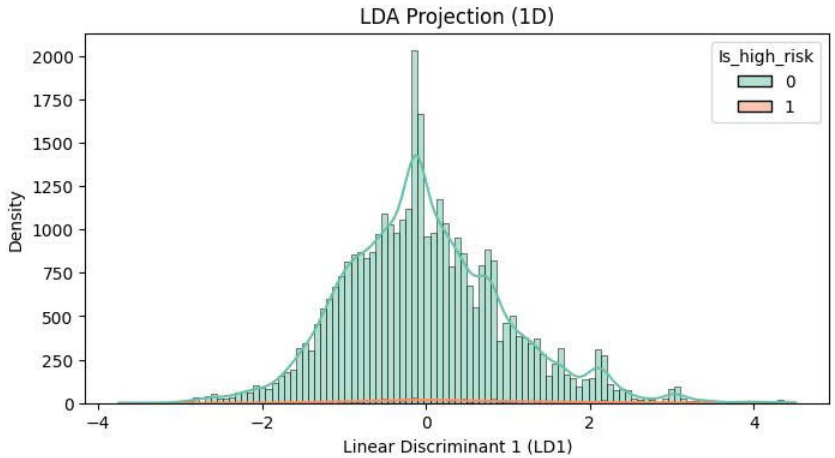


Hình 2.5: Biểu đồ phân tán PCA.

Phân tích biệt thức tuyến tính (LDA)

Kết quả LDA

- Giảm xuống 1 chiều (LD1) để tối đa hóa khoảng cách giữa hai lớp.
- **Ưu điểm**: Tìm ra hướng chiếu có khả năng phân tách tốt hơn PCA.
- **Hạn chế**: Vùng chồng lấn (overlap) giữa hai lớp vẫn rất lớn.



Hình 2.7: Phân phối mật độ trên trục LDA.

Nội dung

Phương pháp tiếp cận

Mục tiêu

- Khám phá cấu trúc ngầm định của dữ liệu.
- Tìm kiếm các nhóm khách hàng có đặc điểm tương đồng mà không cần nhãn.
- Sử dụng dữ liệu đã giảm chiều (PCA) để tăng tốc độ và giảm nhiễu.

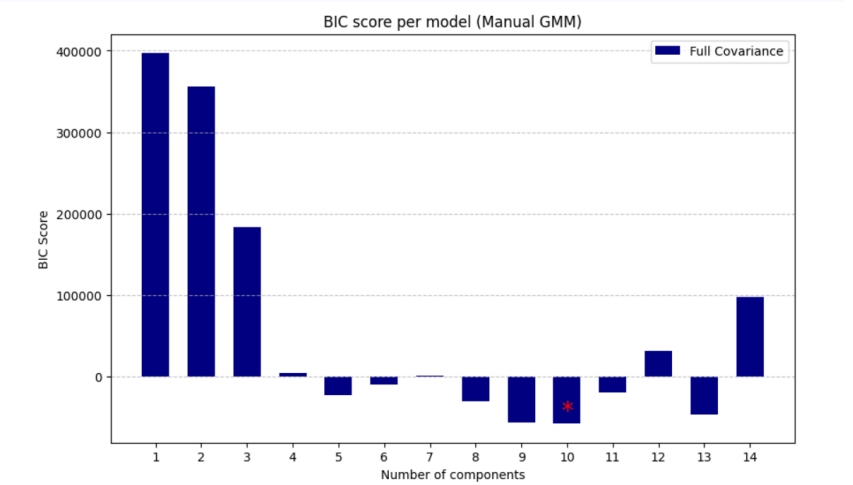
Các thuật toán sử dụng

1. **Gaussian Mixture Model (GMM)**: Phân cụm dựa trên xác suất (phân phối chuẩn).
2. **DBSCAN**: Phân cụm dựa trên mật độ, xử lý tốt nhiễu.

Thực nghiệm GMM

Lựa chọn số cụm tối ưu

- Sử dụng chỉ số **BIC (Bayesian Information Criterion)**.
- Giá trị BIC thấp nhất tại **10 thành phần** (clusters).
- Đầu vào: Dữ liệu sau giảm chiều PCA.

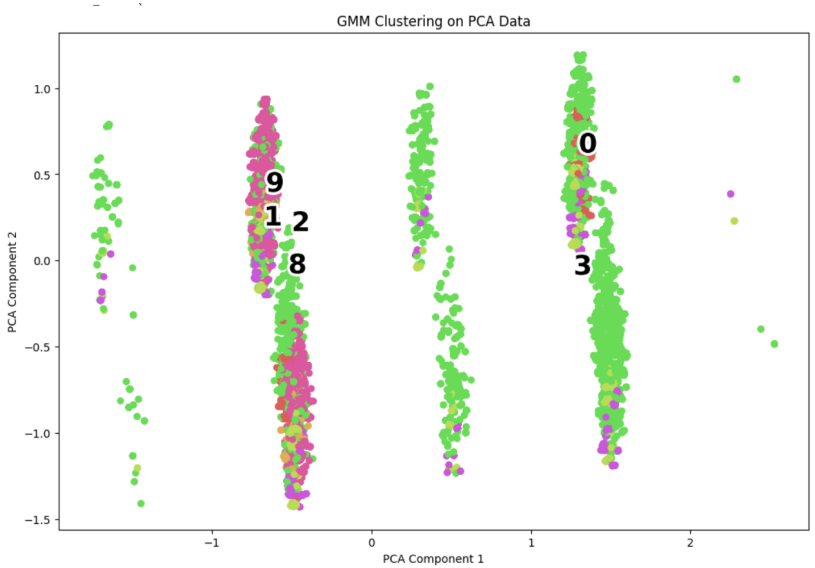


Hình 9: Chỉ số BIC cho từng mô hình

Kết quả phân cụm GMM

Đánh giá định lượng

- **Silhouette Score**: 0.009 (Rất thấp) $\rightarrow$ Các cụm chồng lấn nhiều.
- **Davies-Bouldin**: 3.58 (Cao) $\rightarrow$ Phân tách kém.
- **Modularity**: 0.736 (Cao) $\rightarrow$ Cấu trúc đồ thị quan hệ tốt hơn cấu trúc hình học.



Hình 10: Kết quả phân cụm GMM trên không gian PCA

GMM: Quan hệ với rủi ro tín dụng

Cụm	Mean (Risk)	Std	Số lượng mẫu
Cluster 0	0.011	0.104	1628
Cluster 1	0.017	0.132	1016
Cluster 3	0.019	0.139	9844
...	...	...	...

Bảng: Thống kê rủi ro theo cụm GMM

Nhận xét

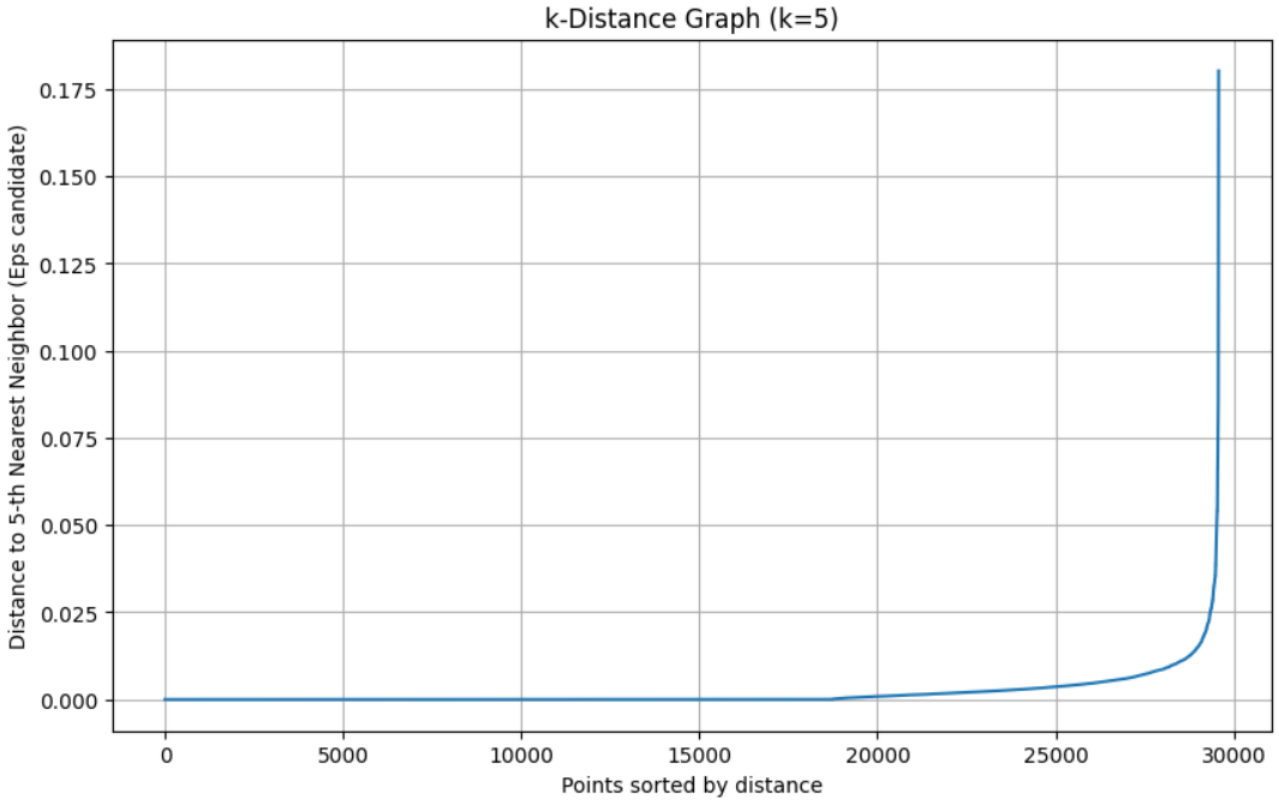
- Tỷ lệ rủi ro giữa các cụm dao động nhỏ (0.011 - 0.020).
- GMM chưa tách biệt rõ ràng được nhóm khách hàng rủi ro cao.
- Phân cụm mang tính mô tả đặc điểm hơn là phân loại rủi ro.

Thực nghiệm DBSCAN

Tham số và Cấu hình

- **Dữ liệu**: 2 thành phần chính đầu tiên của PCA (2D).
- **Chọn tham số**: Dựa trên biểu đồ k-distance ($k = 5$).
- **Ưu điểm**: Tự động phát hiện số cụm và loại bỏ điểm nhiễu (outliers).

Thực nghiệm DBSCAN (2)

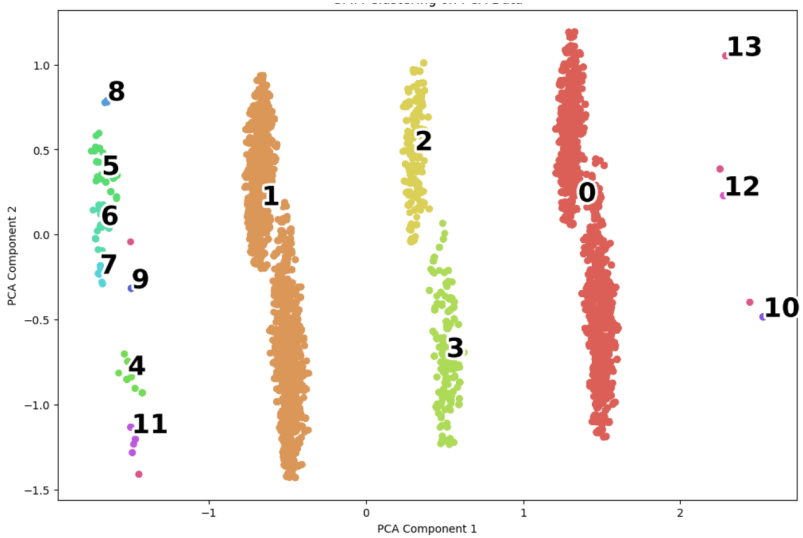


Hình 11: Biểu đồ k-distance xác định Epsilon

Kết quả phân cụm DBSCAN

Chỉ số đánh giá

- **Silhouette Score**: 0.2895 (Khá hơn GMM).
- **Davies-Bouldin**: 1.85 (Tốt hơn GMM).
- Các cụm có sự tách biệt nhất định trong không gian đặc trưng.



Hình 12: Kết quả phân cụm DBSCAN

DBSCAN: Phát hiện nhóm rủi ro cao

Cluster	Mean (Risk)	Count	Đặc điểm
Cluster 0	0.0176	8526	Nhóm lớn, rủi ro thấp
Cluster 1	0.0157	19495	Nhóm lớn, rủi ro thấp
Cluster 11	**0.0833**	**12**	**Nhóm nhỏ, rủi ro cao**
Cluster 8	**0.0769**	**13**	**Nhóm nhỏ, rủi ro cao**

Kết luận quan trọng

- **Xu hướng nghịch biến**: Cụm kích thước càng nhỏ, tỷ lệ rủi ro càng cao.
- DBSCAN hiệu quả trong việc cô lập các nhóm "thiểu số" (High Risk) thành các cụm riêng biệt.
- Hỗ trợ tốt cho việc khoanh vùng khách hàng đáng ngờ.

Tổng kết phần Phân cụm

- **GMM**: Phù hợp để tìm hiểu cấu trúc tổng quát, nhưng yếu trong việc tách biệt rủi ro tín dụng do sự chồng lấn dữ liệu lớn.
- **DBSCAN**: Hiệu quả hơn hẳn trong bài toán này. Khả năng phát hiện nhiễu và gom nhóm dựa trên mật độ giúp tách được các nhóm khách hàng rủi ro cao (thường là các điểm dị biệt/nhóm nhỏ) ra khỏi đa số an toàn.

Nội dung

Mô hình Multilayer Perceptron (MLP)

Kiến trúc mạng

- **Input**: 27 đặc trưng (sau One-Hot).
- **Hidden Layers**: 2 lớp (64 node và 32 node), hàm kích hoạt ReLU.
- **Output**: 1 node (Sigmoid) dự đoán xác suất rủi ro.

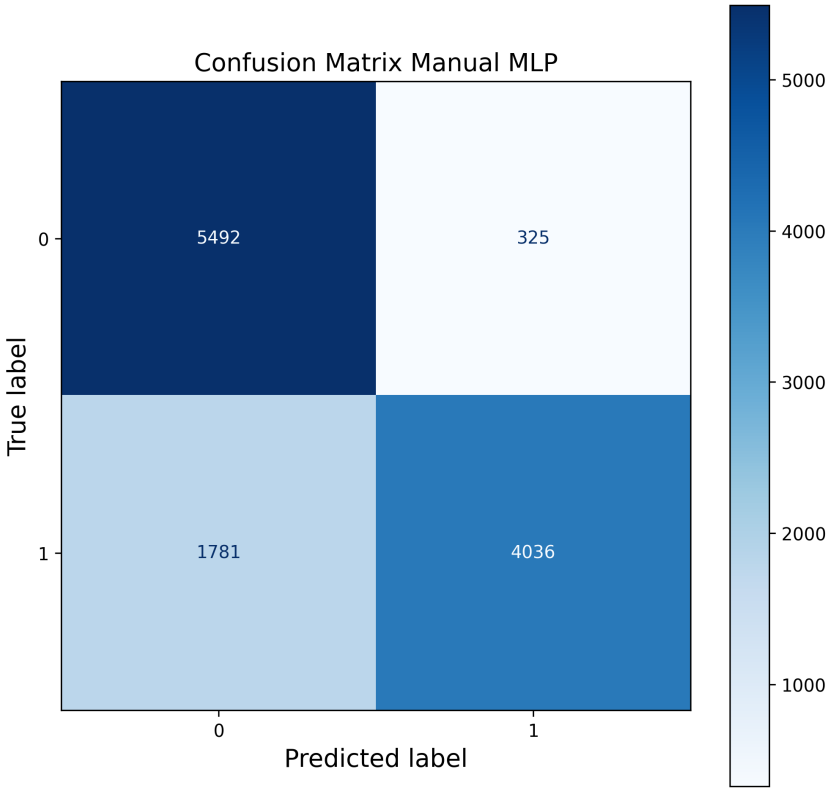
Kết quả mô hình MLP

Dữ liệu	Tỷ lệ	Accuracy	Precision (1)	Recall (1)	F1 (1)
Gốc	**8:2**	**0.82**	**0.93**	**0.69**	**0.79**
Gốc	7:3	0.78	0.95	0.59	0.72
PCA	-	0.82	0.88	0.75	0.81
LDA	-	0.56	0.54	0.79	0.64

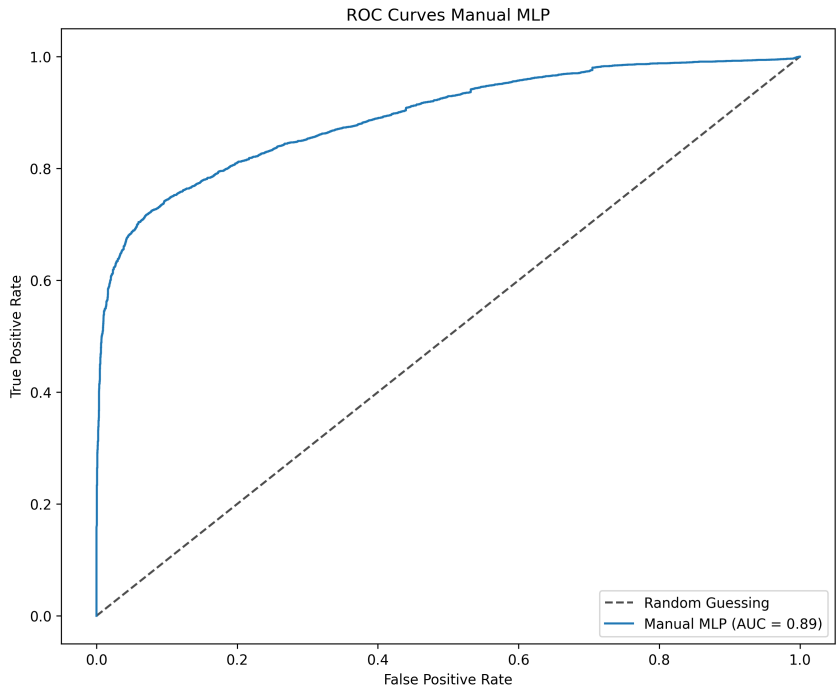
Bảng: Hiệu năng MLP trên các cấu hình

- **Accuracy 0.82** là kết quả tốt trên dữ liệu gốc.
- PCA duy trì hiệu suất tốt (0.82), trong khi LDA thất bại (0.56) do dữ liệu phi tuyến.

Đánh giá MLP: ROC và Confusion Matrix



Hình 13: Ma trận nhầm lẫn MLP



Hình 14: ROC Curve (AUC = 0.89)

Kết luận: MLP xử lý tốt quan hệ phi tuyến, AUC 0.89 rất cao.

Mô hình Gradient Boosting

Nguyên lý Cấu hình

- **Ensemble Learning**: Kết hợp 100 cây quyết định (weak learners) tuần tự.
- **Tham số**: Learning rate $= 0.1$, Max depth $= 3$.
- **Mục tiêu**: Tối ưu hàm Log-Loss cho phân loại nhị phân.

Kết quả Gradient Boosting

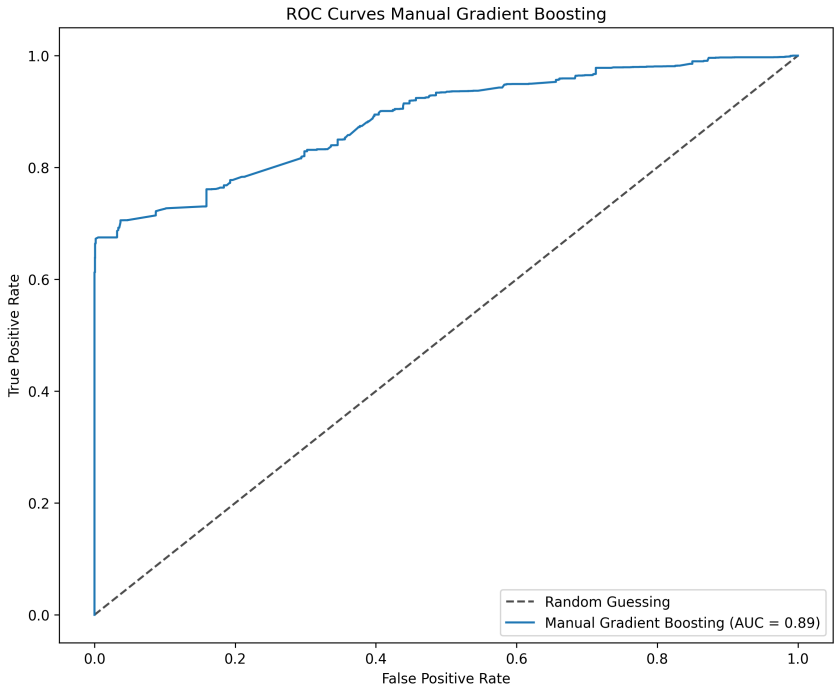
Kịch bản	Accuracy	F1 (Lớp 1)
Gốc (8:2)	0.78	0.78
Gốc (6:4)	0.85	0.82
PCA	0.66	0.69

Bảng: Hiệu năng Gradient Boosting

Nhận xét

- Tỷ lệ 8:2 cho kết quả cân bằng nhất.
- **Vấn đề với PCA**: Hiệu năng giảm sâu (0.66). Lý do: Cây quyết định cắt vuông góc trục, PCA xoay trục làm mất tính trực quan này.

Gradient Boosting: ROC và Đánh giá



Hình 16: ROC Gradient Boosting (AUC = 0.89)

Kết luận

Mô hình có khả năng học cực tốt (AUC 0.89), nhưng không phù hợp khi kết hợp với giảm chiều PCA.

Mô hình SVM (RBF Kernel)

Cấu hình

- Sử dụng **Kernel RBF** để xử lý dữ liệu phi tuyến.
- **C = 10.0**: Ưu tiên phân loại đúng dữ liệu huấn luyện (chấp nhận margin nhỏ hơn).
- Dữ liệu được cân bằng bằng SMOTE.

Kết quả SVM

Dữ liệu	Accuracy	Precision (1)	Recall (1)	F1 (1)
Gốc (8:2)	**0.86**	**0.87**	**0.86**	**0.86**
PCA	0.81	0.82	0.80	0.81
LDA	0.56	0.54	0.80	0.64

Bảng: Hiệu năng SVM

- **Độ chính xác cao nhất (0.86)** và rất cân bằng giữa hai lớp.
- PCA hoạt động tốt với SVM (0.81), là giải pháp thay thế tốt để giảm chi phí tính toán.

Tổng kết hiệu năng mô hình

Kết luận cuối cùng

- **SVM (RBF)** là mô hình tốt nhất với Accuracy 0.86 và F1-score 0.86.
- **MLP** và **Gradient Boosting** đều đạt AUC 0.89, rất mạnh mẽ.
- **Khuyến nghị**: Sử dụng SVM hoặc MLP trên dữ liệu gốc (hoặc PCA) để xây dựng hệ thống phê duyệt tín dụng.
- **Lưu ý**: Tránh dùng LDA cho dữ liệu này vì cấu trúc phi tuyến phức tạp.

Nội dung

Tại sao chuyển sang Hồi quy?

Ý tưởng chủ đạo

- Bài toán gốc: Phân loại nhị phân (0 - An toàn, 1 - Rủi ro) $\rightarrow$ Quyết định cứng nhắc.
- Hạn chế: Không phản ánh được "mức độ" rủi ro khác nhau giữa các khách hàng cùng nhóm.
- **Giải pháp**: Chuyển sang dự đoán giá trị liên tục (Regression).

Mục tiêu

- Ước lượng "Điểm tín dụng" hoặc xác suất rủi ro cụ thể.
- Giá trị dự đoán càng cao $\rightarrow$ Rủi ro càng lớn.
- Cung cấp thông tin chi tiết hơn cho ngưỡng chấp nhận tín dụng của ngân hàng.

Mô hình XGBoost Regressor

Nguyên lý hoạt động

- Là thuật toán Boosting: Kết hợp nhiều cây quyết định yếu theo tuần tự.
- Hàm dự đoán là tổng có trọng số của các cây: $\hat{y}_i = \sum f_t(x_i)$.
- Tối ưu hóa bằng khai triển Taylor bậc hai (Gradient + Hessian) giúp hội tụ nhanh.

Khác biệt so với XGBoost Phân loại

- **Hàm mất mát**: Sử dụng MSE (Mean Squared Error) thay vì Log-Loss.
- **Đầu ra**: Mỗi lá cây chứa một giá trị thực (score) đóng góp trực tiếp vào kết quả, thay vì log-odds.

Ưu điểm của XGBoost Regressor

- **Độ chính xác cao**: Nhờ cơ chế boosting sửa sai tuần tự và khả năng bắt các mối quan hệ phi tuyến.
- **Kiểm soát Overfitting**: Tích hợp sẵn các thành phần điều chuẩn ($L1/L2$) và tham số γ (độ phức tạp cây).
- **Hiệu suất**: Xử lý song song và tối ưu hóa bộ nhớ đệm.

XGBoost Regressor đặc biệt hiệu quả khi cần một thước đo rủi ro liên tục và chính xác.

Mô hình Random Forest Regressor

Nguyên lý "Trí tuệ đám đông"

- Xây dựng tập hợp B cây quyết định độc lập (Bagging).
- Kết quả dự đoán là **trung bình cộng** của các cây thành phần:

$$\hat{y}_i = \frac{1}{B} \sum_{b=1}^{B} T_b(x_i)$$

Kỹ thuật ngẫu nhiên hóa

1. **Bootstrapping**: Mỗi cây học trên một tập mẫu con lấy ngẫu nhiên có hoàn lại.
2. **Feature Randomness**: Tại mỗi nút, chỉ xét một tập con các đặc trưng ngẫu nhiên để tách nhánh.

Tổng kết phần Hồi quy

XGBoost Regressor	**Random Forest Regressor**
Mô hình tuần tự (Boosting)	Mô hình song song (Bagging)
Độ chính xác cực cao	Độ ổn định cao, chống nhiễu tốt
Cần tinh chỉnh tham số kỹ	Dễ sử dụng, ít tham số hơn
Phù hợp để tối ưu hóa điểm số	Phù hợp để đánh giá đặc trưng

Kết luận: Việc mở rộng sang hồi quy giúp ngân hàng có công cụ định lượng rủi ro linh hoạt hơn thay vì chỉ chấp nhận/từ chối.