# Báo cáo: HOME CREDIT - CREDIT RISK MODEL STABILITY

## INTRODUCTION

Notebook "CatBoost, LightGBM Ensemble" trên Kaggle, được tạo bởi PXCAI666, cung cấp một phân tích chi tiết và triển khai các kỹ thuật tổ hợp sử dụng các mô hình CatBoost và LightGBM cho một cuộc thi học máy.

## Chuẩn bị Dữ liệu

Notebook bắt đầu với các bước tiền xử lý dữ liệu, bao gồm xử lý các giá trị thiếu, mã hóa các đặc trưng phân loại và chuẩn hóa các đặc trưng số. Tác giả đảm bảo rằng dữ liệu được làm sạch và sẵn sàng cho việc mô hình hóa bằng cách thực hiện phân tích dữ liệu khám phá và trực quan hóa các thống kê chính.

## Kỹ thuật Xây dựng Đặc trưng

Xây dựng đặc trưng là một phần không thể thiếu của notebook. Các kỹ thuật chính bao gồm:

- Tạo các đặc trưng mới dựa trên kiến thức miền.

- Tạo các đặc trưng đa thức và tương tác để nắm bắt các mối quan hệ phức tạp.

- Các đặc trưng thời gian để tận dụng thông tin liên quan đến thời gian.

## Cập nhật số liệu

- Ngày 7/5/2024 , nhóm tôi tìm kiếm được một notebook của PXCAI666 trong HOME CREDIT với số điểm 0.585, và chúng tôi quyết định đây là baseline của đồ án.

- Ngày 9/5/2024 , sau khi đã cập nhật lại P1-max\_depth = 20 , P2-max\_depth = 16 , P1-num\_leaves = 64 , P2-num\_leaves = 50 , AUC = 1 thì nhận được số điểm là 0.585 , sau đó lại nhận ra khi AUC đạt giá trị 1, điều này cho thấy mô hình phân loại hoàn hảo trên tập kiểm tra, điều này thường là một dấu hiệu đáng nghi ngờ. Trong thực tế, việc đạt được AUC = 1 là rất hiếm và có thể chỉ ra rằng mô hình đã bị quá khớp (overfitting), ngoài ra P1,P2 còn khá cao khiến mô hình càng trở nên phức tạp nên tiếp tục điều chỉnh.

- Ngày 12/5/2024 , sau khi cập nhật P1 = 10 , P2 = 8 , AUC = 0 và cho learning rate = 0.03 thì nhận được số điểm khả quan hơn là 0.591 , việc đó cho thấy mô hình đã được cải thiện phần nào , tuy nhiên learning rate còn thấy chưa đủ cao để khiến cho mô hình hoàn hảo hơn nên tiếp tục chỉnh sửa.

- 23/5/2024 , cập nhật learning rate lên 0.05 thì nhận được số điểm là 0.593, đây là điểm tốt nhất hiện tại của nhóm em và vẫn còn đang thấy mô hình này có thể train 1 cách hoàn hảo hơn nữa.

## Mô Hình Cơ Bản

Hai mô hình chính được sử dụng:

- \*\*CatBoost\*\*: Một triển khai của gradient boosting trên cây quyết định đặc biệt hiệu quả với dữ liệu phân loại.

- \*\*LightGBM\*\*: Một khung làm việc gradient boosting rất hiệu quả sử dụng các thuật toán học cây.

## Tổ hợp

Notebook nhấn mạnh sức mạnh của việc tổ hợp mô hình để cải thiện hiệu suất dự đoán. Các kỹ thuật được thảo luận bao gồm:

- \*\*Stacking\*\*: Điều này bao gồm huấn luyện một meta-model trên các dự đoán của các mô hình cơ bản. Các mô hình cấp đầu tiên (CatBoost và LightGBM) tạo ra các dự đoán, sau đó được sử dụng làm đầu vào cho một mô hình cấp thứ hai để đưa ra dự đoán cuối cùng.

- \*\*Blending\*\*: Tương tự như stacking nhưng đơn giản hơn, blending sử dụng một tập giữ lại để tạo ra các dự đoán từ các mô hình cơ bản, sau đó được trung bình hóa hoặc kết hợp bằng một mô hình khác.

## Đánh giá

Hiệu suất mô hình được đánh giá bằng các chỉ số như ROC-AUC. Notebook cung cấp một so sánh chi tiết về hiệu suất của các mô hình cá nhân và mô hình tổ hợp, nhấn mạnh sự cải thiện đạt được thông qua tổ hợp.

## Tầm Quan Trọng của Đặc trưng

Notebook phân tích tầm quan trọng của các đặc trưng khác nhau, cho thấy các biến nào đóng góp nhiều nhất vào sức mạnh dự đoán của các mô hình. Phân tích này giúp hiểu rõ hành vi của mô hình và cung cấp thông tin chi tiết về các mẫu dữ liệu cơ bản.

## Kết luận

Notebook kết luận với những thông tin chi tiết về các đặc trưng ảnh hưởng nhất và hiệu quả của phương pháp tổ hợp. Nó cũng gợi ý các cải tiến tiềm năng và các thử nghiệm bổ sung có thể được thực hiện.

Tóm lại, kỹ thuật xây dựng và lựa chọn đặc trưng là những bước quan trọng nhất trong cuộc thi này, và một mô hình đơn lẻ được tối ưu hóa cao có thể vượt trội so với các tổ hợp phức tạp.