ĐƠN VỊ TỔ CHỨC

ĐỐI TÁC CHIẾN LƯỢC

CUỘC THI

DATA SCIENCE TALENT COMPETITION 2024

VÒNG 2

Đội thi: Vô Gia Cư

Ngày thực hiện: 30/9 - 6/10/2024

Mục lục

1. Giới thiệu
2. Xử lí & lựa chọn dữ liệu
3. Dashboard trực quan hóa dữ liệu
4. Kết luận & phát hiện bổ sung
5. Giải pháp

1. Giới thiệu

Tổng quan

- Trong lĩnh vực tài chính và ngân hàng, **dữ liệu lịch sử tín dụng** đóng vai trò quan trọng trong việc đánh giá khả năng tín dụng của khách hàng.
- Đội thi sẽ phân tích và xử lí dữ liệu lịch sử tín dụng của khách hàng. Dựa trên các biểu đồ và phân tích trực quan, các bạn sẽ giúp ngân hàng hiểu hơn về đặc điểm của khách hàng có khả năng trả nợ đúng hạn và khách hàng không có khả năng trả nợ đúng hạn.

1. Giới thiệu

Bộ dữ liệu

- Bộ dữ liệu cung cấp các thông tin chi tiết bao gồm:
  - **ID cá nhân của khách hàng.**
  - **Các biến lịch sử tín dụng:** Lịch sử vay nợ, khoản vay hiện tại, dư nợ, số lần quá hạn trả nợ và mức độ quá hạn, các khoản vay có bảo đảm hoặc không bảo đảm, số lần tra cứu thông tin tín dụng *(chi tiết xem hình 1).*
  - **Label:** Label: 1 - khách hàng bị quá hạn (default), 0 - khách hàng trả nợ đúng hạn.

Nền tảng mô hình

- Ngôn ngữ sử dụng: Python.
- Nền tảng sử dụng: Google Colab, PowerBI.

Hình 1. Bảng mô tả đặc trưng

Tên cột/ nhóm cột	Mô tả
customer_id	Mã định danh của khách hàng
label	0: Khách hàng trả nợ đúng hạn 1: Khách hàng quá hạn trả nợ
COUNT	Số khoản vay theo từng loại ngắn hạn - trung hạn - dài hạn từ ngân hàng - tổ chức phi ngân hàng
NUMBER_OF_LOANS_	Tổng tất cả khoản vay (không phân biệt loại) từ ngân hàng - tổ chức phi ngân hàng
NUMBER_OF_CREDIT_CARDS_	Tổng số thẻ tín dụng được cấp bởi ngân hàng - tổ chức phi ngân hàng
NUMBER_OF_RELATIONSHIP_	Tổng số mối quan hệ tài chính mà khách hàng có với ngân hàng - tổ chức phi ngân hàng
NUM_NEW_LOAN_TAKEN_xM	Tổng số khoản vay mới từ ngân hàng - tổ chức phi ngân hàng mà khách hàng thực hiện trong 3 - 6 - 9 - 12 tháng trước đó
OUTSTANDING_BAL_LOAN_CURRENT	Số dư nợ của các khoản vay tính đến hiện tại
OUTSTANDING_BAL_LOAN_xM	Số dư nợ của các khoản vay trong 3 - 6 - 9 - 12 tháng trước đó
OUTSTANDING_BAL_CC_xM	Số dư nợ của thẻ tín dụng trong 3 - 6 - 9 - 12 tháng trước đó
OUTSTANDING_BAL_ALL_xM	Tổng số dư nợ cho tất cả các sản phẩm tài chính trong 3 - 6 - 9 - 12 tháng trước đó
OUTSTANDING_BAL_LOAN_xM_yM	Số chênh lệch giữa các số dư nợ của các khoản vay trong hai khoảng thời gian
OUTSTANDING_BAL_CC_xM_yM	Số chênh lệch giữa các số dư nợ của thẻ tín dụng trong hai khoảng thời gian
OUTSTANDING_BAL_ALL_xM_yM	Số chênh lệch giữa tổng các số dư nợ cho tất cả các sản phẩm tài chính trong hai khoảng thời gian
INCREASING_BAL_xM_	Số tăng lên trong số dư nợ của các khoản vay - thẻ tín dụng - tất cả sản phẩm tài chính trong 3 - 6 tháng
OUTSTANDING_BAL_CC_CURRENT	Số dư nợ của thẻ tín dụng tính đến hiện tại
CREDIT_CARD_MONTH_SINCE_xDPD	Số tháng kể từ khi khách hàng có thanh toán quá hạn 10 - 30 - 60 - 90 ngày trên khoản thanh toán thẻ tín dụng gần nhất
CREDIT_CARD_NUMBER_OF_LATE_PAYMENT	Tổng số lần thanh toán trễ của khách hàng trên các thẻ tín dụng
ENQUIRIES_FROM_FOR_xM	Số lượt tra cứu tín dụng liên quan đến các khoản vay - thẻ tín dụng từ ngân hàng - tổ chức phi ngân hàng trong 3 - 6 - 9 - 12 tháng trước đó
ENQUIRIES_FROM_xM_yM	Số lượt tra cứu tín dụng liên quan đến các sản phẩm tài chính từ ngân hàng - tổ chức phi ngân hàng trong 2 khoảng thời gian
OUTSTANDING_BAL_ALL_CURRENT	Tổng số dư nợ cho tất cả các sản phẩm tài chính tính đến hiện tại

2. Xử lí & lựa chọn dữ liệu

Làm sạch dữ liệu

Các dữ liệu BTC cung cấp là nguồn dữ liệu thô, cần qua quá trình xử lý trước khi đưa vào phân tích.
Một số vấn đề trong nguồn dữ liệu (có thể do yếu tố khách quan hoặc chủ quan):

- **Giá trị ô dữ liệu bị trống ('NaN'):** Mỗi cột đều có tới 10% ô trống (2000/20000 ô).
- **Lỗi sai logic** (ví dụ: _COUNT_BANK + _COUNT_NON_BANK = _COUNT + 1).

Từ đó, đội thi đã đưa ra các giải pháp để xử lý dữ liệu như sau:

- Điều chỉnh hàng loạt giá trị các cột để đảm bảo điều kiện logic.

Ví dụ: giảm 1 đơn vị các ô trong cột phần liên quan đến COUNT, nhằm dẫn đến các logic phù hợp: _COUNT_BANK + _COUNT_NON_BANK = _COUNT, _LOANS = SHORT_ + MID_ + LONG_.

- Điền các ô giá trị trống, theo các logic đã nghiên cứu trên. Hoặc theo sự tương quan với các giá trị cùng cột.
- Xóa đi các hàng không đủ dữ liệu sau quá trình trên. Điều này đảm bảo, do số lượng hàng xóa đi chỉ đạt khoảng 1-2% so với số hàng ban đầu (20000 hàng).

2. Xử lí & lựa chọn dữ liệu

Lựa chọn dữ liệu để phân tích

Các dữ liệu sau khi được làm sạch, sẽ cần lựa chọn các khía cạnh phù hợp để trực quan hóa và xây dựng giải pháp sau này.

- Loại bỏ những đặc trưng có sự trùng lặp cao, ví dụ: *CREDIT_CARD_MONTH_* đều có giá trị là 431 và 'NaN' ở các ô dữ liệu trong cột -> Không có ý nghĩa về sự tương quan.
- Đội thi dựa vào các tài liệu tham khảo từ lĩnh vực tài chính ngân hàng cùng những đánh giá trực quan, quyết định lựa chọn các đặc trưng để tiến hành trực quan hóa như sau: *Short Term Count, Mid Term Count, Long Term Count, Number of Loans, Number of Credit Card, Number of Relationship, Outstanding BAL CC, Outstanding BAL Loan, Outstanding BAL All.*
- Đội thi sẽ tiến hành trực quan hóa các đặc trưng nêu trên, sau đó phân tích và đánh giá nhằm mục đích lựa chọn những đặc trưng quan trọng nhất, ảnh hưởng lớn tới việc khách hàng có trả nợ đúng hạn hay không.

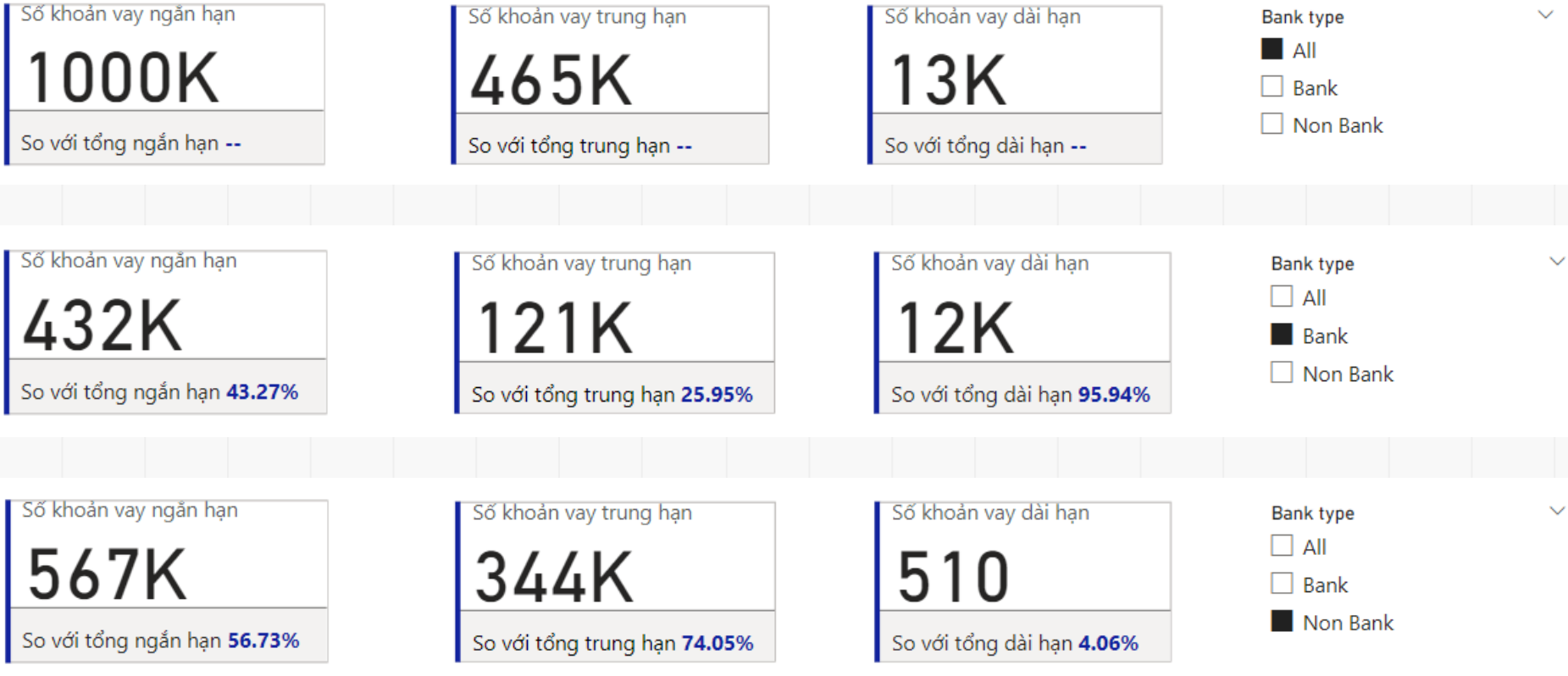
3. Dashboard trực quan hoá dữ liệu

- Công cụ trực quan : Power BI.

- Tập nguồn dữ liệu: Dữ liệu của ban tổ chức sau khi đã được tiền xử lý.
- Các cột sử dụng: Short Term Count, Mid Term Count, Long Term Count, Number of Loans, Number of Credit Card, Number of Relationship, Outstanding BAL CC, Outstanding BAL Loan, Outstanding BAL All.
- Các loại biểu đồ sử dụng : Card, Pie Chart, Stacked Columns chart, Clustered Columns chart.

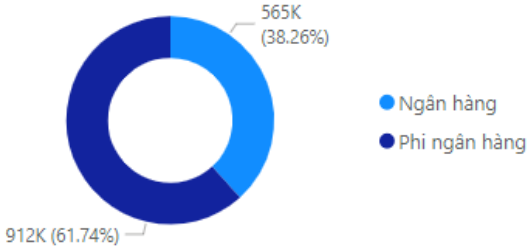
Link nhóm đã trực quan dữ liệu qua PowerBI

Một số hình ảnh dữ liệu trực quan



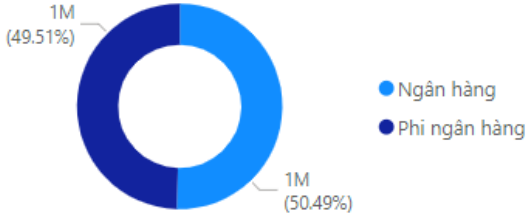
Một số hình ảnh dữ liệu trực quan

Khoản vay từ ngân hàng and Khoản vay từ phi ngân hàng

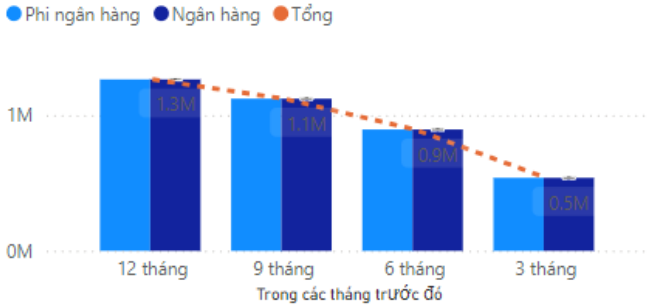


Số thẻ tín dụng cấp từ ngân hàng và phi ngân hàng

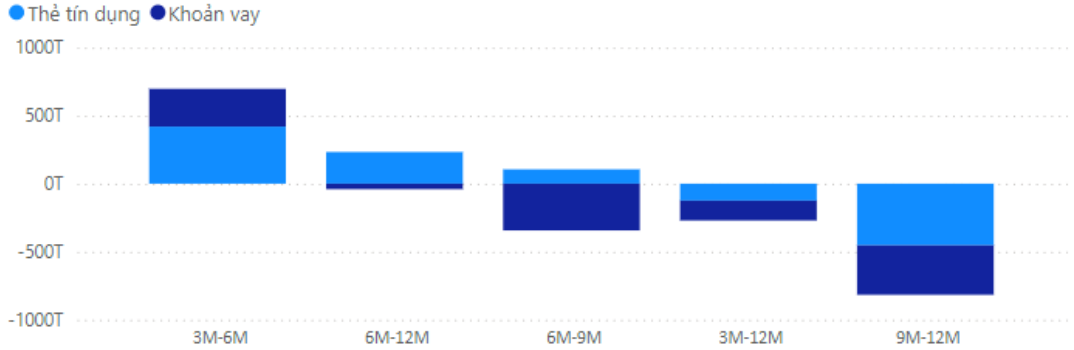
Số mối quan hệ tài chính từ ngân hàng và phi ngân hàng



Khoản vay mới từ ngân hàng và phi ngân hàng



Sự chênh lệch của số dư nợ trong khoảng thời gian



4. Kết luận & phát hiện bổ sung

- Với các khoản vay ngắn hạn và trung hạn, khách hàng có xu hướng sẽ vay từ các tổ chức phi ngân hàng do thủ tục nhanh chóng, đơn giản và tính linh hoạt trong điều khoản. Ngược lại, các khoản vay dài hơn thường sẽ ở tổ chức ngân hàng do lãi suất thấp và độ an toàn cao.

- Khoản vay và số thẻ tín dụng được cấp từ ngân hàng đều lớn hơn các tổ chức phi ngân hàng

-> Sự tin cậy và sự an tâm của khách hàng vào các tổ chức ngân hàng

-> Các thẻ tín dụng từ các ngân hàng thường có sự kết nối với các chính sách ưu đãi,….

- Xu hướng chung là khoản vay giảm dần khi thời gian thu hẹp lại, đặc biệt rõ ở giai đoạn 3 tháng.

-> Các khoản vay ngắn hạn ít phổ biến hơn do khách hàng không đủ khả năng trả nhanh, lãi suất cho vay ngắn hạn có thể cao hơn. -> Không an toàn.

- Có sự sụt giảm đáng kể trong số dư nợ ở các giai đoạn từ 3M-6M và 6M-9M.

-> Khách hàng có khả năng thanh toán nhanh hoặc đây là dấu hiệu của tình trạng thanh toán nhanh trong các kỳ hạn này hoặc là dấu hiệu của một chiến lược vay - tái vay ngắn hạn.

5. Giải pháp

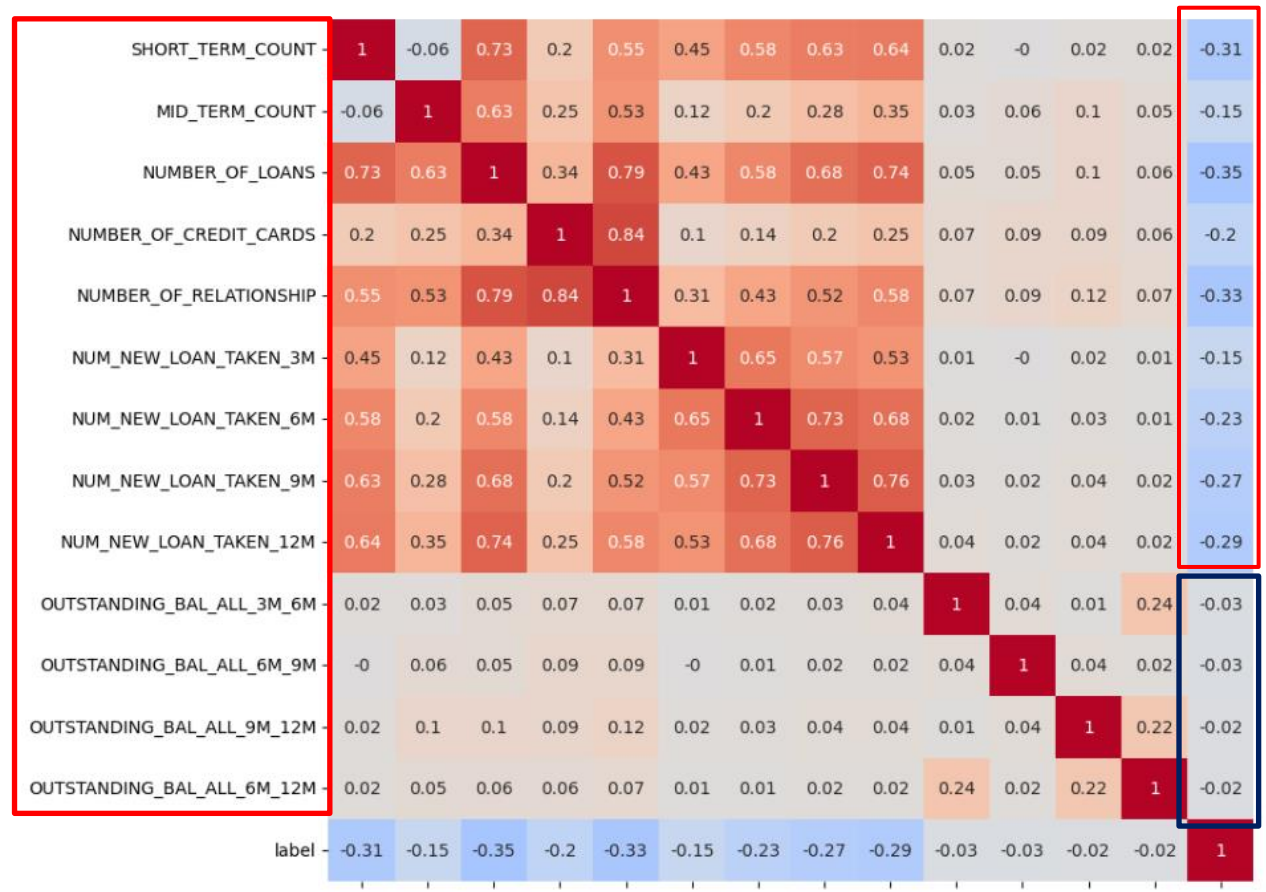
Lựa chọn giải pháp hợp lý & giải thích

Đội thi sau quá trình tìm hiểu đã đưa ra giải pháp **sử dụng mô hình học máy** để dự báo khả năng thanh toán của khách hàng dựa trên các dữ liệu đã trực quan hóa.

- Đội thi dựa vào **biểu đồ trực quan** và **bảng chỉ số tương quan** (sns.heatmap()) của thư viện seaborn (*chi tiết xem ở hình 2*) để lựa chọn ra các đặc trưng quan trọng nhất.
- **Bảng chỉ số tương quan** biểu diễn chỉ số tương quan của *label* với các đặc trưng khác. **Giá trị 1 thể hiện 2 đặc trưng tỉ lệ thuận** còn **-1 thể hiện tỉ lệ nghịch**.
- Các đặc trưng càng tỉ lệ với *label* (giá trị tương quan càng cao) càng cho thấy mức độ quan trọng của đặc trưng đó với việc *label* là 0 hay 1. Tuy nhiên việc lựa chọn quá nhiều đặc trưng có tương quan cao có thể khiến mô hình học máy gặp phải tình trạng overfitting.
- Vì vậy, đội thi đã chọn ra được 13 đặc trưng quan trọng nhất (trong đó có 4 đặc trưng có tương quan thấp nhưng mang nhiều ý nghĩa khi xét trên thực tế) để huấn luyện và kiểm tra mô hình (*chi tiết xem ở hình 2*).

Hình 2. Bảng chỉ số tương quan

13 đặc trưng được đội thi lựa chọn



Các đặc trưng có tương quan cao

Các đặc trưng tuy tương quan không cao nhưng quan trọng khi xét trên thực tế

5. Giải pháp

Lựa chọn giải pháp hợp lý & giải thích

- Sau khi lựa chọn được 13 đặc trưng quan trọng, đội thi chạy thử một số mô hình học máy xây dựng dựa trên dữ liệu được cung cấp và đưa ra đánh giá về các mô hình như sau:

Model	Accuracy	Recall (macro avg)	Precision (macro avg)	F1-Score (macro avg)
Logistic Regression	0.882386	0.683422	0.925610	0.734034
Decision Tree	0.860543	0.712571	0.778899	0.737474
Random Forest	0.856623	0.706626	0.770124	0.730482
XGBoost	0.878465	0.706419	0.851928	0.748511
Naive Bayes	0.468776	0.625589	0.584974	0.456903

5. Giải pháp

- Đội thi đề cao chỉ số **Accuracy** và **Recall** của mô hình (Recall được tính bằng số dự đoán khách hàng bị quá hạn của mô hình trên tổng số khách hàng thực sự quá hạn). Vì vậy, các mô hình có chỉ số Recall cao sẽ giúp đưa ra những cảnh báo giúp quán triệt triệt để những khách hàng có khả năng trả nợ muộn.
- Qua thử nghiệm và phân tích, đội thi lựa chọn được 3 mô hình học máy hiệu quả cho việc dự đoán khách hàng bị quá hạn là: **Decision Tree**, **Random Forest** và **XGBoost**.

Model	Accuracy	Recall (macro avg)	Precision (macro avg)	F1-Score (macro avg)
Logistic Regression	0.882386	0.683422	0.925610	0.734034
Decision Tree	0.860543	0.712571	0.778899	0.737474
Random Forest	0.856623	0.706626	0.770124	0.730482
XGBoost	0.878465	0.706419	0.851928	0.748511
Naive Bayes	0.468776	0.625589	0.584974	0.456903

Lời cảm ơn

Trước tiên đội thi xin được cảm ơn ban tổ chức, ban giám đã dành thời gian để đọc bản báo cáo của nhóm chúng tôi. Bản báo cáo này là thành quả của một quá trình phân tích, tìm hiểu cũng như tham khảo các nguồn tài liệu, các kiến thức về trực quan hóa, lập trình cùng các vấn đề liên quan. Sau đó được nghiên cứu và sử dụng các kiến thức, công cụ đã nghiên cứu để xây dựng những biểu đồ trực quan và mô hình học máy hiệu quả cho đề bài.

Xin cảm ơn ban tổ chức của cuộc thi vì đã tạo ra một sân chơi bổ ích cho chúng tôi được trải nghiệm, giao lưu và phát triển những kỹ năng quan trọng.

Xin chân thành cảm ơn!