TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN MÔN HỌC

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

BANK MARKETING

NHÓM I

Nguyễn Trường Lâu – 17520676 (Ad)

Trần Công Minh – 17520763

Bùi Đức Cường – 17520301

GVGD: Ths. Nguyễn Thị Anh Thư

MỤC LỤC

I.TÌM HIỂU DỮ LIỆU

+ Nguồn: tạo bởi Paulo Cortez và Sérgio Moro

(Paulo Cortez)

(Sérgio Moro)

+ Hiện là Tiến sĩ trong Data Mining (đại học Minho và Lisbon Bồ Đào Nha).

+ Dữ liệu liên quan đến các chiến dịch tiếp thị trực tiếp từ xa (thông qua điện thoại) của một tổ chức ngân hàng Bồ Đào Nha.

+ Mục tiêu phân loại: để dự đoán xem khách hàng có đăng ký gửi tiền có kỳ hạn hay không?

+ Mục đích phân tích dữ liệu:

*Tăng hiệu quả của chiến dịch tiếp thị từ xa của ngân hàng.

*Phân tích khách hàng để dự đoán hành vi tiết kiệm của khách hàng, xác định khách hàng tiềm năng

II.CHUẨN BỊ DỮ LIỆU

1.LÀM SẠCH DỮ LIỆU

+ Điền giá trị thiếu: Không có dữ liệu thiếu

+ Khử dữ liệu nhiễu:

a. Balance (Số dư trung bình hằng năm)

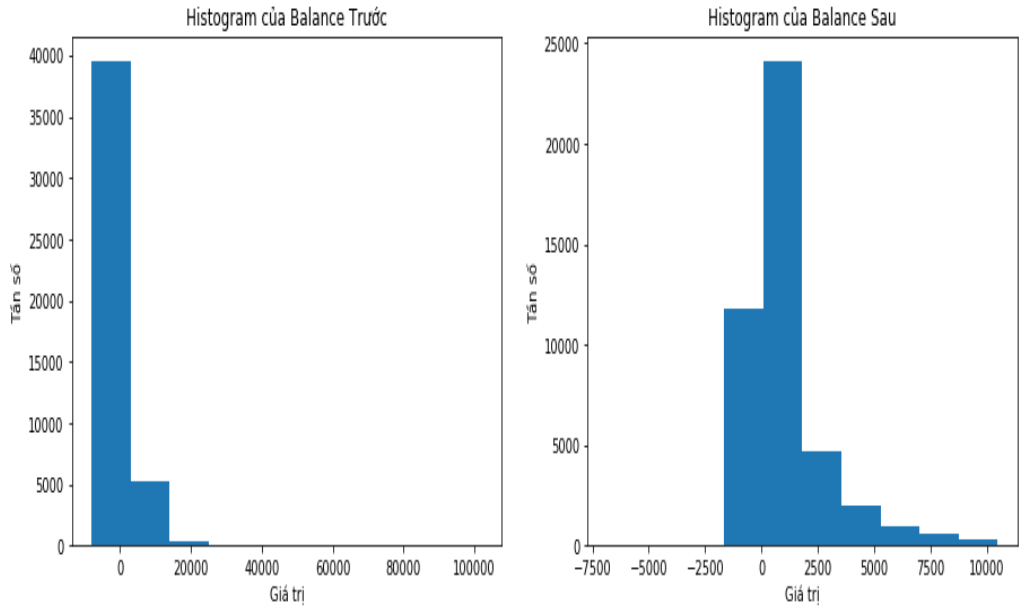
b. Duration (thời gian liên lạc)

a. KHỬ NHIỄU TRÊN BALANCE

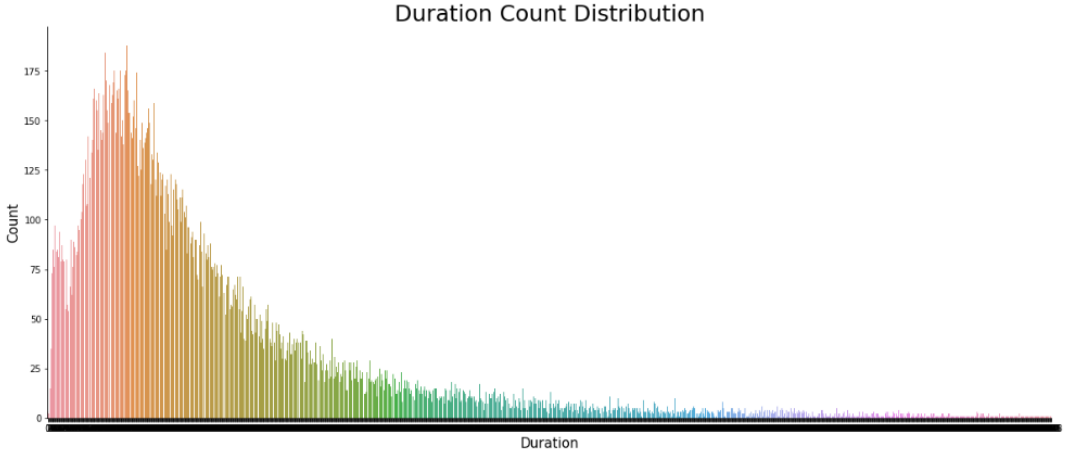
+ Chuẩn hóa dữ liệu: chuẩn hóa theo Z-score:

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

+ Loại bỏ nhiễu: loại bỏ các điểm dữ liệu vượt ngưỡng -3->3, sau bước này dữ liệu giảm xuống 1811 điểm dữ liệu.



b. KHỬ NHIỄU TRÊN DURATION



+ Chuẩn hóa dữ liệu: Đơn vị ban đầu được tính bằng giây -> chuyển sang phút

+ Khử dữ liệu nhiễu:

* Thống kê cho thấy: các điểm dữ liệu <5s và lớn hơn 65 phút đều cho y='no'.

* Thời gian trao đổi là 0s thì y='no' là hợp lý do đó có thể xem đây là dữ liệu nhiễu.

* Loại bỏ các điểm dữ liệu có thời gian <5s và lớn hơn 65 phút.

2. Mã hóa dữ liệu

Mã hóa chuỗi thành số theo thứ tự Alpha:

VD: Thuộc tính Marital (tình trạng hôn nhân): divorced ,married, single -> mã hóa thành 0 - 1 – 2.

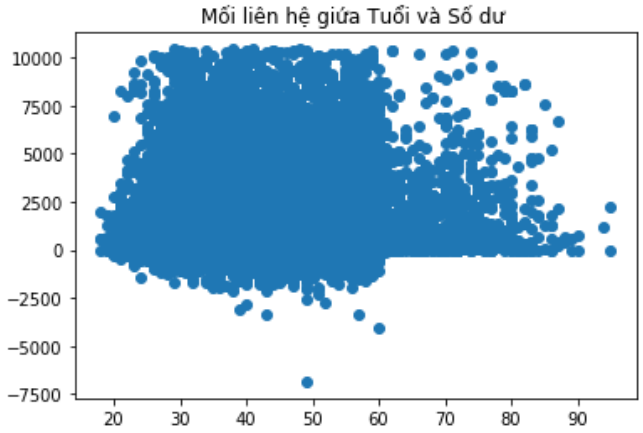
3. RÚT GỌN DỮ LIỆU

+ Tại trường dữ liệu: contact (thiết bị liên lạc giữa khách hàng và ngân hàng): unknown, telephone (điện thoại bàn), cellular (di động).

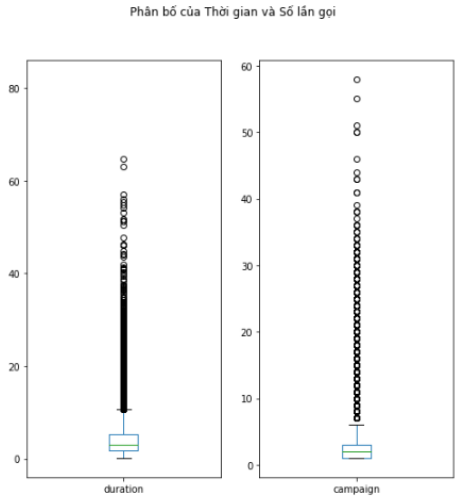
+ Trường dữ liệu này không ảnh hưởng đến kết quả tham gia đăng ký gửi tiền có kỳ hạn

=> Xóa thuộc tính Contact. Data sẽ còn lại 15 thuộc tính.

III. PHÂN TÍCH DỮ LIỆU

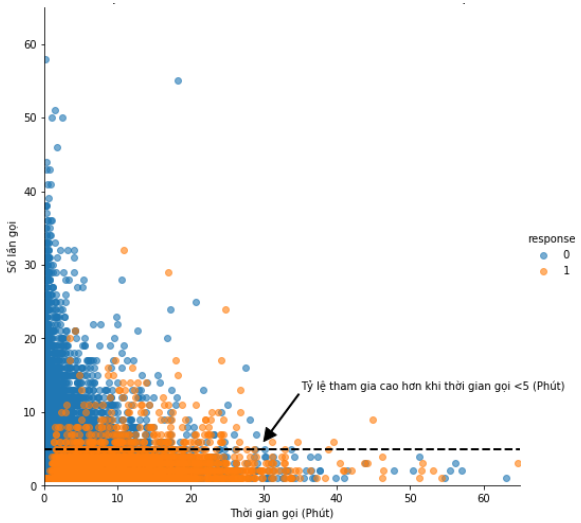


+ Khi trên 60 tuổi thì số dư hằng năm dường như không âm và nhỏ hơn 5000 USD -> đây là tuổi nghỉ hưu, nguồn cung chủ yếu là lương hưu.



+ Hầu hết khách hàng có thời gian trao đổi với ngân hàng từ 5->6 phút và từ 2->3 lần.

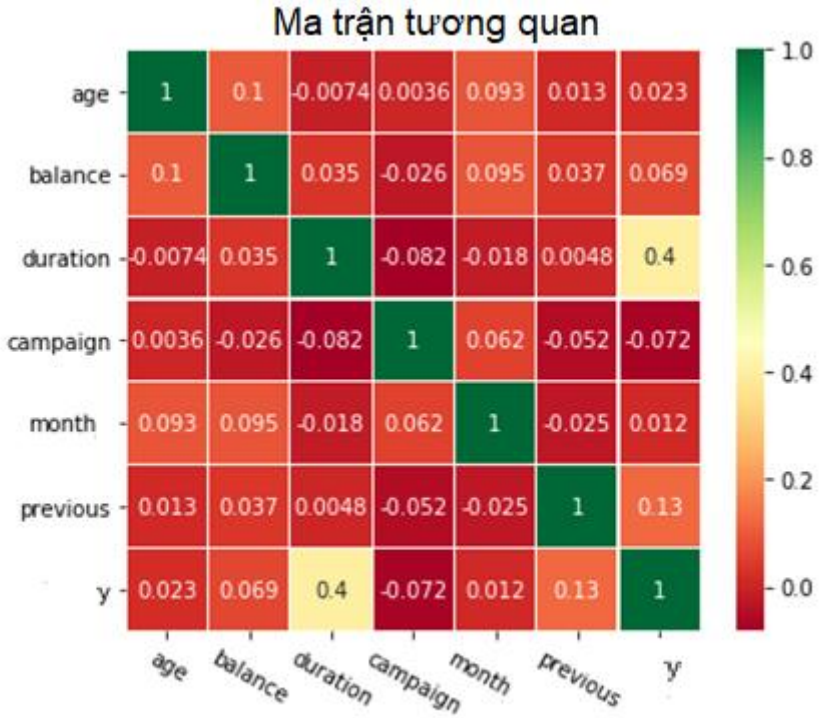
+ Tuy nhiên vẫn có những khách hàng có thời gian trao đổi lên tới gần 60 phút hay có số lần trao đổi lên đến 60 phút.



+ Khi thời gian gọi từ 3->10 phút thì hầu hết các khách hàng sẽ đăng ký tham gia gửi tiền có kỳ hạn.

+ Tuy nhiên khi số lần gọi càng nhiều thì khả năng khách hàng tham gia càng thấp => đồng nghĩa với việc khách hàng đang cảm thấy phiền hà khi bị ngân hàng làm phiền nhiều lần => không tham gia.

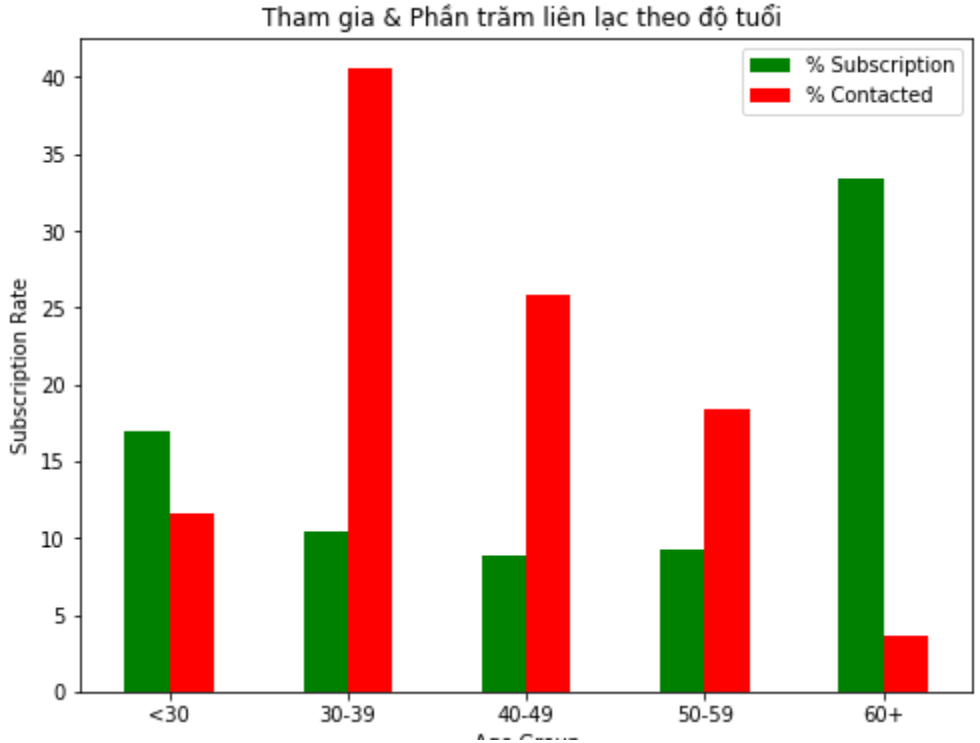
=> Ngân hàng cần điều chỉnh lại số lần gọi và thời gian của các cuộc gọi sao cho hợp lý, và lý tưởng nhất là: số lần gọi là từ 2->5 lần và thời gian cho mỗi lần gọi là từ 3->15 phút.



+ Dữ liệu đa phần tương quan thuận với nhau.

+ Các dữ liệu tương quan lớn đến y là: duration, previous, month, balance.

+ Nhận xét: do dữ liệu đa phần tương quan đến nhau nên có thể phương pháp phân loại bằng Native Bayes sẽ không hiệu quả.

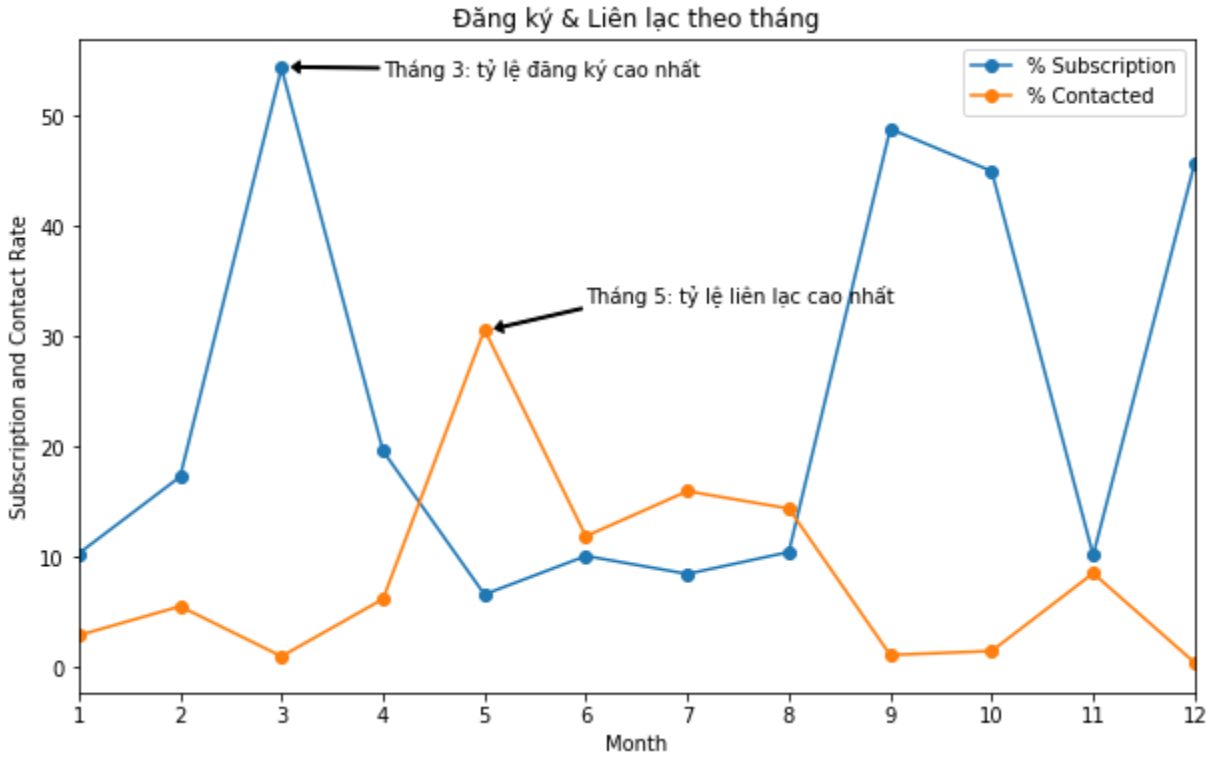


+ Ta nhận thấy độ tuổi đăng ký nhiều nhất là từ 60 trở lên và đứng thứ hai là độ tuổi từ 18->29 tuổi.

+ Được giải thích: ngoài 60 ta có xu hướng tiết kiệm hơn sau khi nghỉ hưu so với trung niên, họ có xu hướng tích cực hơn với mục tiêu là tạo thư nhập đầu tư cao. Tiền gửi có kỳ hạn là đầu tư ít rủi ro nên được người lớn tuổi ưu tiên.

+ Đối với người trẻ: họ không có đủ tiền và chuyên môn để tham gia vào các khoản đầu tư lớn nên gửi tiền có kỳ hạn giúp họ có một thu nhập ổn định hơn.

+ Tuy nhiên các thanh màu đỏ cho thấy ngân hàng đang tập trung nỗ lực tập trung tiếp thị vào nhóm trung niên. Vì thế để nâng cao hiệu quả tiếp thị thì ngân hàng nên chuyển hướng tiếp thị cho các khách hàng trẻ và già hơn trong tương lai.

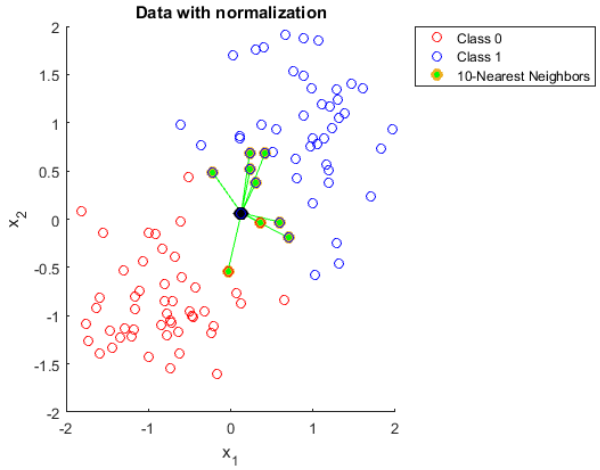


+ Ngân hàng liên hệ với khách hàng hầu hết là tháng tháng 5 -> tháng 8, tuy nhiên tỷ lệ đăng ký lại có xu hướng đi ngược lại với tỷ lệ liên lạc.

+ Tỷ lệ đăng ký ở tháng 2->4 và 9->10 lại cao hơn trong khi tỷ lệ liên hệ lại thấp hơn => cho thấy ngân hàng đang đi sai đường, ngân hàng nên chuyển thời gian tiếp thị vào mùa xuân và mùa thu. Tuy nhiên ngân hàng nên cận trọng vì yếu tố này mang tính chất thời gian nên có thể thay đổi theo từng năm.

IV. MÔ HÌNH PHÂN LỚP DỮ LIỆU

1.K – NEAREST NEIGHBORS



+ Hãy cho tôi biết bạn của bạn là ai, tôi sẽ nói bạn là người như thế nào .

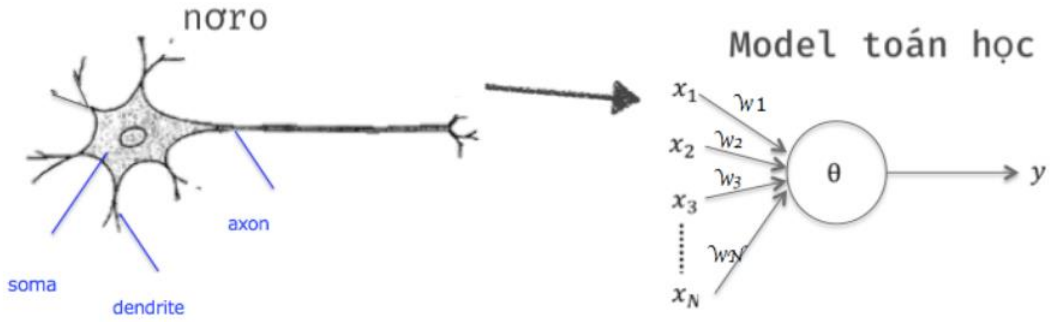
+ Một mẫu mới được gán vào lớp có nhiều mẫu giống với nó nhất trong số k mẫu gần nhất

+ Thử trên nhiều K khác nhau, chọn K cho ta độ chính xác là cao nhất thì ta chọn.

+ Vẽ biểu đồ cho ta thấy sự biến đổi của độ chính xác theo từng K, khi đó ta sẽ chọn được K “tốt nhất”.

+ Khi K=5 thì độ chính xác đạt cao nhất, Acc=88.9%

2. DEEP NEURAL NETWORK



+ Mô hình mô phỏng nơ-ron trong hệ thống thần kinh con người. Theo tính chất truyền tín hiệu: khi neuron nhận tín hiệu đầu vào từ các dendrite, khi tín hiệu vượt qua một ngưỡng(threshold) thì tín hiệu sẽ được truyền đi sang neuron khác (Neurons Fire) theo sợi trục(axon).

+ Neural của model toán học ở đây cũng được mô phỏng tương tự:

$$y = a(w_1x_1 + w_2x_2 + w_3x_3 - \theta)$$

+ Giải thích:

* Khi ta làm việc, não chúng ta tiếp nhận rất nhiều thông tin từ các giác quan khác nhau.

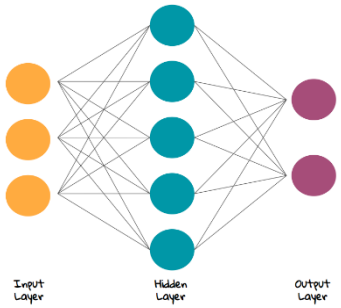
* Như vậy làm sao để não ta phân biệt được thông tin quan trọng hay không, đó là một cấu trúc phức tạp của não bộ cho phép nó gắn các trọng số vào mỗi thông tin.

+ Công thức cho mỗi node:

$$y = a(w_1x_1 + w_2x_2 + w_3x_3 + b)$$

Trong đó: a là hàm kích hoạt có nhiệm vụ chuẩn hóa dữ liệu, b được gọi là bias.

+ Gộp nhiểu Unit ta được một mạng:



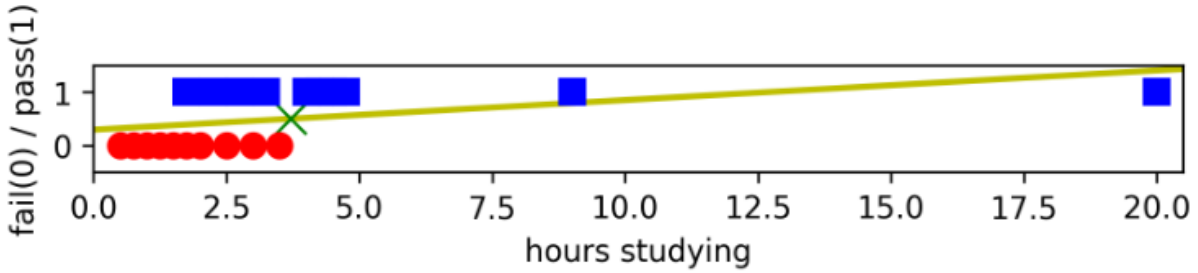
+ Sau nhiều lần Feedforward và Backprogation thì ta học được bộ tham số W phù hợp.

+ Mỗi mỗi điểm dữ liệu đầu vào, thì ta dự đoán được nó thuộc lớp nào dựa vào Layer Output. Nếu tại Layer Output, giá trị Unit nào lớn hơn thì điểm dữ liệu đó sẽ thuộc lớp đó.

3. LOGISTIC REGRESSION

+ Linear Regression:

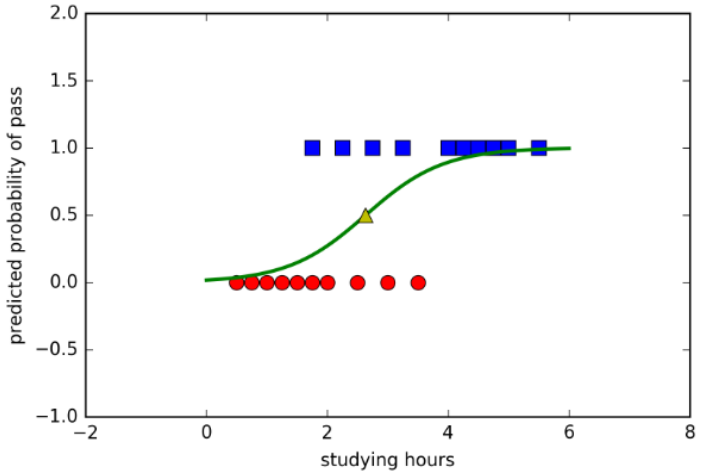
$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$$



+ Đầu ra dự đoán của Logistic regression:

$$f(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x})$$

+ Trong đó θ được gọi là logistic function.



+ Việc học của chúng ta sẽ là học ra bộ trọng số W cho hàm trên sao cho độ lỗi là tối ưu nhất.

+ Đầu ra dự đoán là xác suất điểm dữ liệu đó thuộc một lớp thõa:

* Nếu y-pre >= @ thì cho dữ liệu đó thuộc lớp thứ nhất.

* Nếu y-pre < @ thì cho dữ liệu đó thuộc lớp thứ hai.

V. ĐÁNH GIÁ MÔ HÌNH

+ Trong 2 lớp, lớp quan trọng hơn sẽ được gọi là Positive (tích cực), lớp còn lại ít quan trọng hơn sẽ là Negative (tiêu cực).

+ Yes = Positive, No = Negative

+ Dự đoán nhầm còn hơn bỏ sót: dự đoán một người có khả năng đăng ký tiền gửi có kỳ hạn sẽ tốt hơn là chúng ta bỏ qua một cơ hội tốt để tăng tiền gửi chô ngân hàng.

Confusion matrix:

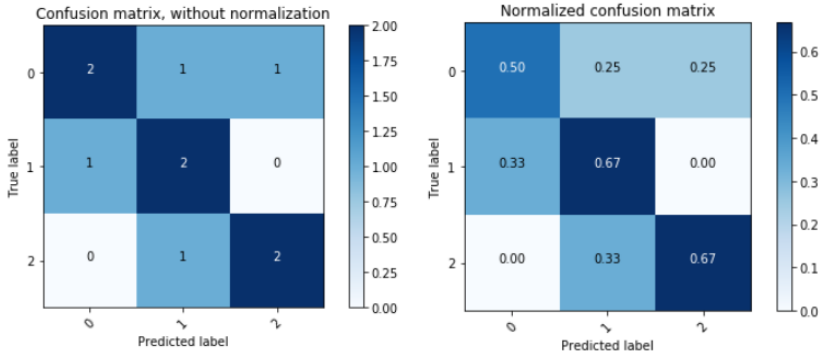
+ Lớp nào được phân loại đúng nhiều nhất

+ Dữ liệu thuộc lớp nào thường bị phân loại nhầm vào lớp khác

Total: 10	Predicted as: 0	Predicted as: 1	Predicted as: 2	
True: 0	2	1	1	4
True: 1	1	2	0	3
True: 2	0	1	2	3

+ Độ lớn của ma trận dựa vào số lớp, Vd khi ta có 3 lớp thì ma trận nhầm lẫn của ta là 3x3.

+ Giá trị tại hàng i cột j là: số lượng điểm đáng lẽ được phân vào lớp I nhưng lại bị phân vào lớp j.



1.KNN

	precision	recall	f1-score	support
0	0.88	0.98	0.93	5158
1	0.11	0.02	0.04	683
accuracy			0.86	5841
macro avg	0.50	0.50	0.48	5841
weighted avg	0.79	0.86	0.82	5841

2.DNN

	precision	recall	f1-score	support
0	0.88	1.00	0.94	5158
1	0.00	0.00	0.00	683
accuracy			0.88	5841
macro avg	0.44	0.50	0.47	5841
weighted avg	0.78	0.88	0.83	5841

3.LR

	precision	recall	f1-score	support
0	0.91	0.98	0.94	5158
1	0.64	0.31	0.42	683
accuracy			0.90	5841
macro avg	0.77	0.64	0.68	5841
weighted avg	0.88	0.90	0.88	5841

VI. KẾT LUẬN VÀ KHUYẾN NGHỊ

+ KẾT LUẬN:

- Khách hàng tiềm năng có 3 đặc trưng cơ bản:

1. Tuổi <30 hoặc >60
2. Là sinh viên hoặc người đã nghỉ hưu
3. Số dư >5000 USD

-Với 3 mô hình đã được đề cập thì ngân hàng có thể dự đoán phản ứng của khách hàng trước gọi trực tiếp để tiếp thị để giảm chi phí và tăng hiệu quả tiếp thị của ngân hàng.

- Ngân hàng nên phân bổ lại thời gian và nổ lực hơn cho những khách hàng được dự đoán là sẽ đăng ký cao, đồng thời gọi ít lại cho những người không có khả năng gửi tiền có kỳ hạn.

- Dự đoán thời lượng gọi và điều chỉnh kế hoạch tiếp thị để không gây ảnh hưởng đến khách hàng, đồng thời nâng cao sự hài lòng của khách hàng đối với ngân hàng.

+ KHUYẾN NGHỊ:

- Chọn thời điểm thích hợp hơn: mùa xuân và mùa thu là hai mùa có tỷ lệ thành công cao nhất. Ngân hàng cần phân tích nhiều dữ liệu hơn để đảm bảo rằng hiêu ứng theo mùa này không đổi theo thười gian để có hướng tiêp thị đúng đắn trong tương lai.

- Cung cấp dịch vụ tốt hơn cho khách hàng: với các thông tin về khách hàng mà ngân hàng có được, ngân hàng nên ước tính khi nào khách hàng cần đầu tư để đưa ra lời khuyên cho khách hàng, mang đến lợi ích cho cả hai.

--------END--------