TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO CUỐI KỲ MÔN
HỆ THỐNG THƯƠNG MẠI THÔNG MINH

BANK MARKETING DATA SET

Người hướng dẫn: **TS. DƯƠNG HỮU PHÚC**

Người thực hiện: **NGUYỄN HOÀNG QUANG NHẬT - 51800220**

NGUYỄN THỊ HỒNG HƯƠNG - 51800284

ĐINH TIẾN BÌNH - 51800525

Lớp: **18050301, 18050402**

Khóa: **22**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO CUỐI KỲ MÔN
HỆ THỐNG THƯƠNG MẠI THÔNG MINH

BANK MARKETING DATA SET

Người hướng dẫn: **TS. DƯƠNG HỮU PHÚC**

Người thực hiện: **NGUYỄN HOÀNG QUANG NHẬT**

NGUYỄN THỊ HỒNG HƯƠNG

ĐINH TIẾN BÌNH

Lớp: **18050301, 18050402**

Khóa: **22**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021

LỜI CẢM ƠN

Chúng em xin chân thành cảm ơn TS Dương Hữu Phúc đã cung cấp những kiến thức vô cùng quý báu về môn học. Trong suốt quá trình giảng dạy, thầy đã nhiệt tình chỉ bảo chúng em rất nhiều để có thể tiến hành thực hiện bài báo cáo này.

Sau khoảng thời gian học tập, với những kiến thức đã tiếp thu được từ thầy song vẫn còn mặt hạn chế từ phía kiến thức cũng như kĩ năng thực hành nên bài báo cáo này không thể tránh khỏi sai và thiếu sót.

Một lần nữa nhóm chúng em xin cảm ơn thầy và mong nhận được những đóng góp quý giá từ thầy để bài báo cáo được chỉnh chu và hoàn thiện hơn nữa.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Chúng tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của TS Dương Hữu Phúc. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào chúng tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do chúng tôi gây ra trong quá trình thực hiện (nếu có).

TP.Hồ Chí Minh, ngày tháng năm

Tác giả

(kí và ghi rõ họ tên)

Nguyễn Thị Hồng Hương

Đinh Tiến Bình

Nguyễn Hoàng Quang Nhật

PHẦN NHẬN XÉT VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

TP. Hồ Chí Minh, ngày tháng năm

(ký tên và ghi rõ họ tên)

Phần đánh giá của GV chấm bài

TP. Hồ Chí Minh, ngày tháng năm

(ký tên và ghi rõ họ tên)

TÓM TẮT

Ngày nay, dữ liệu chính là vua. Sử dụng dữ liệu với các mục đích khác nhau cho tác dụng to lớn đối với doanh nghiệp thương mại. Ngoài ra, một trong những cách mà tổ chức doanh nghiệp có thể thực hiện khi đứng trên thị trường là nắm bắt và kiểm tra thành thạo thông tin khách hàng để phát triển hơn nữa trải nghiệm của khách hàng. Bài báo cáo này thu thập tập dữ liệu được thu thập từ trang UCI Machine Learning Repository. Thông tin được xác định với các chiến dịch tiếp thị ngân hàng của các cơ sở cuộc gọi. Mục tiêu chính là trực quan hóa dữ liệu trên Tableau và thử nghiệm xem liệu khách hàng có quyết định đăng ký dịch vụ gửi tiền có kỳ hạn hay không. Bên cạnh đó báo cáo có trình bày sơ lược lý thuyết một số thuật toán như Naive Bayes, Decision Tree, Support Vector Machine, Logistic Regression và K-Nearest Neighbor và tiến hành đưa ra các xác suất dự đoán dựa trên tập dữ liệu này.

Mục lục

Danh sách hình vẽ

Chương 1

TỔNG QUAN ĐỀ TÀI

1.1 Giới thiệu đề tài

Một trong những tổ chức ngân hàng Bồ Đào Nha đã thực hiện một chiến dịch tiếp thị dựa trên các cuộc điện thoại từ năm 2008 đến năm 2010. Các hồ sơ về nổ lực trong chiến dịch tiếp thị của họ có sẵn dưới dạng tập dữ liệu. Bài báo cáo này sẽ áp dụng các thuật toán học máy để phân tích tập dữ liệu dự đoán xem khách hàng có đăng ký tiền gửi có hạn không, cũng nhằm tìm ra các chiến thuật hiệu quả nhất giúp ngân hàng trong chiến dịch tiếp theo thuyết phục thêm nhiều khách hàng đăng ký gửi tiền có kỳ hạn. Tập dữ liệu chứa các tính năng phân loại và số khác nhau với 41188 mẫu dữ liệu. Dữ liệu được gắn nhãn. Mục tiêu là dự đoán liệu khách hàng có đăng ký tiền gửi có kỳ hạn hay không. Xử lý trước dữ liệu được thực hiện cùng với phân tích dữ liệu phù hợp. Kết quả của các thuật toán khác nhau sẽ được so sánh.

1.2 Phát biểu bài toán

Trong một môi trường thông tin điện tử phát triển mạnh mẽ như ngày nay thì các tổ chức hàng ngày thiết lập các bộ dữ liệu, tìm hiểu và kết hợp chúng để đưa ra những quyết định hiệu quả để cải thiện thương mại. Việc công nghệ đi kèm với lượng dữ liệu dồi dào đóng vai trò then chốt thì khả năng quyết định nhanh chóng và chuẩn xác là một

lợi thế cạnh tranh vô cùng quan trọng. Hệ thống thương mại thông minh (BI - Business Intelligence) mang đến lợi ích giúp người dùng đưa ra những quyết định đúng đắn, gia tăng giá trị của doanh nghiệp nói chung. Đặc biệt là trong ngành ngân hàng, việc phân tích, xử lý và khai thác những nguồn thông tin khổng lồ có được từ khách hàng giúp đưa ra các chiến lược tiếp thị thu hút khách hàng cũng như tiềm năng dịch vụ hay rủi ro của ngân hàng. Tốc độ thành công của tiếp thị ngân hàng phụ thuộc vào kết quả và sự lựa chọn để đưa ra các dự đoán và chiến lược chính xác hơn.

Từ thực tế đó, trên cơ sở nhu cầu ngày càng cao của việc đưa ra các chiến lược tiếp thị trong ngành ngân hàng thì chúng em đã quyết định chọn đề tài "Bank Marketing Data Set" nhằm nghiên cứu và tìm hiểu về quá trình xử lý tập dữ liệu của một ngân hàng để dự đoán xem khách hàng có đăng ký dịch vụ gửi tiền có kỳ hạn hay không trong chiến dịch tiếp thị này.

1.3 Mục tiêu chọn đề tài

Dự đoán, đưa ra các xác suất, tỉ lệ mà khách hàng tại ngân hàng có quyết định đăng ký gửi một khoản tiền có thời hạn hay không.

- Áp dụng các thuật toán học máy xử lý và phân tích dữ liệu.
- Tìm hiểu thuật toán hồi quy logistics, K-Nearest Neighbors, Support Vector Machine, Decision Tree và Naive Bayes để xác định.
- Trực quan hóa tập dữ liệu trên Tableau.

1.4 Phạm vi đề tài

- Sử dụng 5 thuật toán là KNN, Decision Tree, hồi quy logistic, SVM và Naive Bayes.
- Dùng Tableau để thể hiện đồ thị trực quan.

1.5 Cấu trúc báo cáo

Bài báo cáo gồm có 6 chương:

Chương 1 – Tổng quan đề tài: Chương này giới thiệu đề tài, trình bày đặt vấn đề bài toán dẫn nhập mục tiêu, phạm vi nghiên cứu và cấu trúc của báo cáo.

Chương 2 – Tổng quan giải thuật: Chương này trình bày tổng quan các lý thuyết giải thuật, và thực nghiệm giải thuật đó trên đề tài.

Chương 3 – Dữ liệu thực nghiệm: Chương này trình bày chi tiết về tập dữ liệu sử dụng trong đề tài, trực quan hóa tập dữ liệu đó bằng đồ thị trên Tableau và cách xử lý dữ liệu.

Chương 4 – Thực nghiệm: Chương này trình bày các kết quả với các thuật toán giới thiệu ở chương 2 và đưa ra nhận xét.

Chương 5 – Kết luận: Chương này trình bày những gì đã làm được và chưa làm được so với mục tiêu ban đầu đặt tra và nêu hướng phát triển trong tương lai.

Chương 6 – Tài liệu tham khảo: Chương này tổng hợp các tài liệu tham khảo và dẫn đến các đường dẫn đó.

Chương 2

TỔNG QUAN GIẢI THUẬT

2.1 Thuật toán K-Nearest Neighbors (KNN)

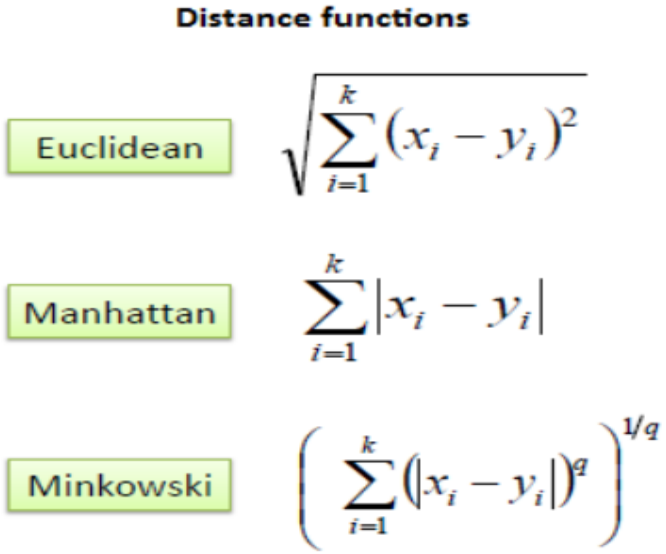
2.1.1 *Cơ sở lý thuyết*

K-nearest neighbor là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning. KNN là thuật toán tìm đầu ra của một điểm dữ liệu mới bằng cách chỉ dựa trên thông tin của K điểm dữ liệu trong training set gần nó nhất (K-lân cận).

- Ứng dụng: Thuật toán KNN có nhiều ứng dụng trong ngành ngân hàng gồm xác định xem khách hàng chậm trả các khoản vay, khách hàng có đăng ký gửi tiền hạn hay không; trong đầu tư như dự đoán phá sản, dự đoán giá cổ phiếu, phân bổ xếp hạng tín dụng trái phiếu doanh nghiệp, tạo ra chỉ số vốn và trái phiếu tùy chỉnh.
- Ưu điểm:
  - Thuật toán đơn giản, dễ dàng triển khai.
  - Độ phức tạp tính toán của quá trình training nhỏ.
  - Việc dự đoán kết quả của dữ liệu mới rất đơn giản.
  - Xử lý tốt với tập dữ liệu bị nhiễu.
- Nhược điểm:

- KNN bị nhiễu dễ đưa ra kết quả không chính xác khi K nhỏ.
- KNN là một thuật toán mà mọi tính toán đều nằm ở khâu test. Việc tính khoảng cách tới từng điểm dữ liệu trong training set sẽ tốn rất nhiều thời gian.
- Độ phức tạp càng tăng khi k càng lớn.

- Ý tưởng thuật toán: Thuật toán KNN cho rằng những dữ liệu tương tự nhau sẽ tồn tại gần nhau trong một không gian, từ đó công việc của ta là sẽ tìm k điểm gần với dữ liệu cần kiểm tra nhất. Việc tìm khoảng cách giữa 2 điểm có nhiều công thức có thể sử dụng, tùy trường hợp mà chúng ta lựa chọn cho phù hợp. Đây là 3 cách cơ bản để tính khoảng cách 2 điểm dữ liệu x, y có k thuộc tính:



Hình 2.1: Một số công thức tính khoảng cách giữa x và y có k thuộc tính

Các bước trong KNN:

1. Ta có D là tập dữ liệu đã được gắn nhãn và A là dữ liệu chưa được phân loại.
2. Đo khoảng cách từ A đến tất cả dữ kiệu đã được phân loại trong D.
3. Chọn K làm khoảng cách nhỏ nhất.

4. Kiểm tra danh sách có khoảng cách ngắn nhất và đếm số lượng lớp xuất hiện.
5. Lấy lớp xuất hiện nhiều nhất.
6. Lớp dữ liệu mới là lớp ở bước 5.

2.1.2 *Thực hành trên tập dữ liệu*

Lấy ra 4 biến độc lập gồm job, marital, housing và loan đã được gắn nhãn phân loại. Chọn ra 10 dữ liệu để xét:

STT	Job_blue-collar (x_1)	Marital_married (x_2)	Housing_yes (x_3)	Loan_yes (x_4)	y
1	1	1	1	0	0
2	0	1	0	0	0
3	0	0	1	0	1
4	0	1	0	0	0
5	0	1	1	0	1
6	0	0	1	0	0
7	1	1	1	0	0
8	1	0	1	0	0
9	0	1	0	0	1
10	1	0	1	0	0
11	**0 (y_1)**	**0 (y_2)**	**0 (y_3)**	**0 (y_4)**	**?**

Hình 2.2: Bảng ví dụ dữ liệu trong tập dữ liệu

Áp dụng hàm Euclidean tính khoảng cách:

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + (x_4 - y_4)^2}$$

STT	Khoảng cách	Xếp hạng
1	1.73	3
2	1	1
3	1	1
4	1	3
5	1.73	1
6	1	1
7	1.73	3
8	1.41	2
9	1	1
10	1.41	2

Hình 2.3: Kết quả tính khoảng cách

Chọn K=5, ta được:

STT	Khoảng cách	Y
2	1	0
3	1	1
5	1	1
6	1	0
9	1	1

Hình 2.4: Kết quả thu được lớp dữ liệu mới

⇒ trong 5 dữ liệu mới được chọn, nhãn 1 chiếm đa số vậy dữ liệu số 11 được phân vào lớp 1.

2.2 Thuật toán hồi quy logistic

2.2.1 *Cơ sở lý thuyết*

Hồi quy từ lâu đã trở thành một phần không thể thiếu trong Data analysis liên quan đến việc tìm hiểu và phân tích mối quan hệ giữa các đối tượng nghiên cứu thể hiện qua biến mục tiêu (biến y) và các biến độc lập (biến giải thích - các biến x).

- **Ứng dụng:**
  - Dự đoán phân loại email có spam hay không
  - Dự đoán khả năng trả nợ của khách hàng.
  - Dự đoán khả năng mua sản phẩm hay là đăng ký dịch vụ,...
- **Phương trình hồi quy logistic:**
  - Probability (p); là xác suất của một biến cố trong một thời gian (p sẽ dao động trong khoảng từ 0 đến 1).
  - odds: là tỷ số giữa xác suất biến cố xảy ra chia cho xác suất biến cố không xảy ra. Odds được xem là một biến liên tục, giá trị của odds không nhất thiết phải nằm tròn khoản từ 0 đến 1 và giá trị của odds bằng 1 khi và chỉ khi p = 0.5. Để tính odds chúng ta áp dụng công thức: $odd = \frac{p}{1-p}$
  - odds ratio: là tỷ số giữa 2 odd vầ được xác định bằng công thứ; $\frac{odd_1}{odd_2} = \frac{P(X,Y)P(\overline{X},\overline{Y})}{P(X,\overline{Y}P(\overline{X},Y)}$

    Ví dụ: *Xét trường hợp 5 người cùng đến xét nghiệm ung thư phổi và kết quả trả về là 1 người trong số họ mắc bệnh.*

    Ta có: $p = \frac{1}{5} = 0.2$

    $$\Rightarrow odds = \frac{p}{1-p} = \frac{0.2}{1-0.2} = 0.25$$
  - logit chính là $\log(odds)$ và được xác định bằng công thức:

    $$logit(p) = \log(odds) = \log(\frac{p}{1-p})$$

- Nếu gọi X là biến tiên lượng và p là xác suất của một biến cố (outcome) thì mô hình hồi quy logistic sẽ được phát biểu như sau:

$$logit(p) = \alpha + \beta X$$

hay:

$$\log(\frac{p}{1-p}) = \alpha + \beta X$$

- Phương trình tổng quát để ước lượng xác suất dạng đa biến như sau:

$\hat{y}$ = Ước lượng P(y = 1 | $x_1, x_2, ..., x_p$) = $\frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}}$ Trong đó:

  - α là hệ số chặn (intercept). Giá trị của z khi tất cả các biến độc lập bằng 0 (X=0).
  - β là hệ số hồi qui (regression cofficients) của các yếu tố nguy cơ (còn gọi là biến độc lập) x1, x2,. . . , xk. Hệ số hồi qui cho biết độ mạnh cũng như chiều của sự ảnh hưởng của các yếu tố nguy cơ đến xác suất xảy ra sự kiện nghiên cứu. Nếu hệ số hồ qui dương thì yếu tố nguy cơ làm tăng khả năng (xác suất) xảy ra của sự kiện nghiên cứu và ngược lại.

2.2.2 *Thực hành trên tập dữ liệu*

- Để dự đoán dữ liệu trong mô hình Logistic Regression thì ta phải tính xác suất để xác định ngưỡng để cho biết dữ liệu đó thuộc nhãn 0 hay 1 thông qua hàm sigmoid.
- Nếu kết quả trả về của hàm Sigmoid đó >= 0.5 thì sẽ là nhãn 1 còn < 0.5 thì sẽ là nhãn 0.
- Ví dụ: một khách hàng có thuộc tính age:40, martial: married, loan:yes, default: no. . . .. thì sẽ được tính qua hàm sigmod có dạng là:

$$Y = \frac{1}{1 + e^{-z}}$$

  - Z= W0 + W1.age + W2.married + W3.loan+ ...

- Giả sử như Y < 0.5 thì ta sẽ kết luẫn khách hàng này thuộc nhãn 0 (y=0) nghĩa là khách hàng này không đăng ký tiền gửi có kỳ hạn.

2.3 Thuật toán Naive Bayes

2.3.1 *Cơ sở lý thuyết*

Naive Bayes Classification (NBC) là một thuật toán phân loại dựa trên tính toán xác suất áp dụng định lý Bayes. Theo định lý Bayes, ta có công thức tính xác suất ngẫu nhiên của sự kiện yy khi biết xx như sau :

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)}$$

Giả sử ta phân chia sự kiện x thành n thành phần khác nhau x1 ,x2 ....,xn. Naive Bayes theo đúng như tên gọi dựa vào một giả thiết x1 ,x2 ....,xn là các thành phần độc lập với nhau. Từ đó ta có thể tính được:

$$P(x \mid y) = P(x_1 \cap x_2 \cap ... \cap x_n \mid y) = P(x_1 \mid y)P(x_2 \mid y)(x_3 \mid y)...P(x_n \mid y)$$

Do đó ta có:

$$P(y \mid x) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

Trong đó $\propto$ là phép tỉ lệ thuận.

- **Các mô hình thuật toán Naive Bayes**:

  - ***Mô hình Bernoulli:*** mô hình này, các feature vector là các giá trị nhị phân 0, 1. Trong đó 1 thể hiện từ có xuất hiện trong dữ liệu , 0 thể hiện từ đó không xuất hiện trong dữ liệu. Xác suất $P(x_i \mid y)$ được tính bằng :

$$P(x_i \mid y) = P(i \mid y).x_i + (1 - P(i \mid y)).(1 - x_i)$$

  Với $P(x_i \mid y)$ tỉ lệ số lần từ x_i xuất hiện trong toàn bộ tập training data có nhãn y.

- ***Mô hình Multinomial:*** Ở mô hình này, các feature vector là các giá trị số tự nhiên mà giá trị thể hiện số lần từ đó xuất hiện trong văn bản. Ta tính xác suất từ xuất hiện trong dữ liệu $P(x_i \mid y)$ như sau:

$$P(x_i \mid y) = \frac{N_i}{N_c}$$

Với N_i là tổng số lần từ x_i xuất hiện trong văn bản, N_c là tổng số lần từ của tất cả các từ X1,....Xn xuất hiện trong dữ liệu.

2.3.2 *Thực hành trên tập dữ liệu*

sex	marital	housing	loan	Default	y
Male	single	yes	no	no	0
Female	divorced	yes	yes	no	1
Male	married	yes	no	no	0
Male	married	yes	no	unknown	0
Female	single	no	no	unknown	1
Male	divorced	no	no	no	0
Male	married	no	yes	no	1
Male	married	yes	no	no	1
Female	married	yes	yes	no	0
Male	divorced	no	no	unknown	0

Hình 2.5: Bảng ví dụ dữ liệu trong tập dữ liệu

P(y=0)=6/10 và P(y=1)=4/10

Ví dụ như cần dư đoán xem khách hàng có các thuộc tính sex:female; marital:married; housing:yes; loan:no; default:unknown để biết người đó có đăng kí tiền gửi có kì hạn hay không thì ta sẽ làm như sau:

Sex	P(0)	P(1)
Male	5/7	2/7
Female	1/3	2/3

Marital	P(0)	P(1)
Single	1/2	1/2
divorced	2/3	1/3
married	3/5	2/5

Housing	P(0)	P(1)
Yes	4/6	2/6
no	2/4	2/4

loan	P(0)	P(1)
Yes	1/3	2/3
no	5/7	2/7

Default	P(0)	P(1)
Unknown	2/3	1/3
no	4/7	3/7

Hình 2.6: Bảng tính toán dự đoán P

- Tính xác suất với P(y=1):
  P(1)*P(female|1)*P(married|1)*P(yes|1)*P(no|1)*(unknown|1)
  =4/10*2/3*2/5*2/6*2/7*1/3 = 0.003
- Tính xác suất với P(y=0):
  P(0)*P(female|0)*P(married|0)*P(yes|0)*P(no|0)*(unknown|0)
  =6/10*1/3*3/5*4/6*5/7*2/3 = 0.038

⇒ Ta thấy 0.038 > 0.003 do đó khách hàng này sẽ có y = 0, nghĩa là khách hàng này không đăng kí tiền gửi có kỳ hạn.

2.4 Thuật toán Support Vector Machine (SVM)

2.4.1 *Cơ sở lý thuyết*

SVM là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại. Trong thuật toán này, chúng ta

vẽ đồ thị dữ liệu là các điểm trong n chiều với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay" (hyper-plane) phân chia các lớp. Hyper-plane nó chỉ hiểu đơn giản là 1 đường thẳng có thể phân chia các lớp ra thành hai phần riêng biệt.

Margin trong SVM là khoảng cách giữa siêu phẳng đến 2 điểm dữ liệu gần nhất tương ứng với các phân lớp. Ví dụ quả táo quả lê đặt trên mặt bàn, margin chính là khoảng cách giữa cây que và hai quả táo và lê gần nó nhất. Điều quan trọng ở đây đó là phương pháp SVM luôn cố gắng cực đại hóa margin này, từ đó thu được một siêu phẳng tạo khoảng cách xa nhất so với 2 quả táo và lê. Nhờ vậy, SVM có thể giảm thiểu việc phân lớp sai (misclassification) đối với điểm dữ liệu mới đưa vào.

Phương thức Kernel:

- Ý tưởng:
  - Định nghĩa K: X x Y $\rightarrow \mathbb{R}$ được gọi là kernel với:

$$\Phi(x).\Phi(y) = K(x,y)$$

  - K thường được hiểu là đại lượng đo độ tương đồng.
- Kernel là các hàm cho trước x, y trả về $\Phi(x), \Phi(y)$. Điều này có nghĩa là chúng ta có thể biến đổi các vectơ đặc trưng thành chiều vô hạn và vẫn không có thêm bất kỳ tính toán nào để tính các điểm chấm.

2.4.2 *Thực hành trên tập dữ liệu*

- Để dự đoán được dữ liệu bằng phương pháp SVM thì ta phải giải phương trình W.X + B = 0 để tìm ra được đường phân cách giữa nhãn 0 và nhãn 1 trong y.
- Tiếp theo đó ta sẽ tìm Margin là đường mở rộng ra từ phương trình W.X + B = 0 sao cho nó càng rộng càng tốt. các biên của margin sẽ chạm vào các điểm đầu tiên của 2 nhãn và điểm đó sẽ được gọi là support vector.

- Để kiểm tra xem dữ liệu cần dự đoán thuộc nhãn 0 hay 1 thì ta sẽ tính phương trình W.X + B cho dữ liệu đó nếu cái phương trình W.X + B này lớn hơn hoặc bằng 1 thì nó thuộc nhãn 1 còn bé hơn hoặc bằng -1 thì nó sẽ thuộc nhãn 0.
- Ví dụ: khách hàng có các thuộc tính age: 55, sex:male, martial:single .... thì sau khi tính phương trình W.X + B >= 1 từ đó ta sẽ kết luận khách hàng này có y =1 tức là có đăng ký tiền gửi có kỳ hạn.

2.5 Thuật toán Decision Tree

2.5.1 *Cơ sở lý thuyết*

Cây quyết định (Decision tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượng có thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary) , Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal. Tóm lại, cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các dữ liệu chưa biết.

- Ưu điểm:
  - Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.
  - Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả.
  - Có thể làm việc với cả dữ liệu số và dữ liệu phân loại.
  - Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê.
  - Có khả năng là việc với dữ liệu lớn.
- Nhược điểm:

- Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thạm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.
- Cây quyết định hay gặp vấn đề overfitting.

- **Entropy:** là thuật ngữ thuộc Nhiệt động lực học, là thước đo của sự biến đổi, hỗn loạn hoặc ngẫu nhiên. Năm 1948, Shannon đã mở rộng khái niệm Entropy sang lĩnh vực nghiên cứu, thống kê với một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau x1,x2,...,xn. Giả sử rằng xác suất để x nhận các giá trị này là $pi = p(x = x_i)$. Ta có công thức tổng quát sau:

$$Entropy(S) = \sum_{i=1}^{c} -p_i.log_2.p_i$$

- **Information Gain:** dựa trên sự giảm của hàm Entropy khi tập dữ liệu được phân chia trên một thuộc tính. Để xây dựng một cây quyết định, ta phải tìm tất cả thuộc tính trả về Infomation gain cao nhất. Để xác định các nút trong mô hình cây quyết định, ta thực hiện tính Infomation Gain tại mỗi nút theo trình tự sau:
  - Bước 1: Tính toán hệ số Entropy của biến mục tiêu S có N phần tử với N_c phần tử thuộc lớp c cho trước

$$H(S) = -\sum_{c=1}^{c} .(N_c)/N.log(N_c)/N$$

  - Bước 2: Tính hàm số Entropy tại mỗi thuộc tính: với thuộc tính x, các điểm dữ liệu trong S được chia ra K child node $S_1, S_2, \ldots, S_K$ với số điểm trong mỗi child node lần lượt là $m_1, m_2, \ldots, m_K$, ta có:

$$H(x,S) = \sum_{k=1}^{K} .(m_k)/N.H(S_k)$$

- Bước 3: Chỉ số Gain Information được tính bằng:

$$G(x,S) = H(S) - H(x,S)$$

- Công thức tổng quát:

$$Gain(S,A) = Entropy(S) - \sum_{v \varepsilon Values(A)} \frac{|S_v|}{|S|}.Entropy(S_v)$$

- **Tiêu chuẩn dừng:** Trong thuật toán Decision tree, với phương pháp chia trên, ta sẽ chia mãi các node nếu nó chưa tinh khiết. Như vậy, ta sẽ thu được một tree mà mọi điểm trong tập huấn luyện đều được dự đoán đúng (giả sử rằng không có hai input giống nhau nào cho output khác nhau). Khi đó, cây có thể sẽ rất phức tạp (nhiều node) với nhiều leaf node chỉ có một vài điểm dữ liệu. Như vậy, nhiều khả năng overfitting sẽ xảy ra. Để tránh trường hợp này, ta có thể dừng cây theo một số phương pháp sau đây:
  - Nếu node đó có entropy bằng 0, tức mọi điểm trong node đều thuộc một class.
  - Nếu node đó có số phần tử nhỏ hơn một ngưỡng nào đó. Trong trường hợp này, ta chấp nhận có một số điểm bị phân lớp sai để tránh overfitting. Class cho leaf node này có thể được xác định dựa trên class chiếm đa số trong node.
  - Nếu khoảng cách từ node đó đến root node đạt tới một giá trị nào đó. Việc hạn chế chiều sâu của tree này làm giảm độ phức tạp của tree và phần nào giúp tránh overfitting.
  - Nếu tổng số leaf node vượt quá một ngưỡng nào đó.
  - Nếu việc phân chia node đó không làm giảm entropy quá nhiều (information gain nhỏ hơn một ngưỡng nào đó).

2.5.2 *Thực hành trên tập dữ liệu*

Ta có mẫu dữ liệu với các trường sau để dự đoán khách hàng có đăng ký tiền vào ngân hàng hay không với các thuộc tính sau: Sex : giới tính, Marital: Hôn nhân, Housing: Nhà

ở, Loan : Tình trạng có vay hay không, Contact: Liên lạc, Default: Vỡ nợ, y : Khách hàng có đăng ký gửi tiền hay không.

sex	marital	housing	loan	contact	Default	y
Male	single	yes	no	cellular	no	0
Female	divorced	yes	yes	cellular	no	1
Male	married	yes	no	cellular	no	0
Male	married	yes	no	cellular	unknown	0
Female	single	no	no	telephone	unknown	1
Male	divorced	no	no	telephone	no	0
Male	married	no	yes	cellular	no	1
Male	married	yes	no	cellular	no	0
Female	married	yes	yes	cellular	no	1
Male	divorced	no	no	cellular	unknown	0

Hình 2.7: Bảng ví dụ dữ liệu trong tập dữ liệu

$Entropy([4+,6-]) = -(4/10).log_2(4/10) - (6/10).log_2(6/10) = 0.97$

- Values (sex) = Male, Female

  $S_{Male} = [1+,6-]$

  $S_{Female} = [3+,0]$

  $Gain(S,sex) = Entropy(S) - (7/10).Entropy(S_{Male}) - (3/10).Entropy(S_{Female})$ = 0.97 - (7/10) * 0.59 - (3/10) * 0 = 0.4

- Value (marital) = single, married, divorced

  $S_{single} = [1+,1-]$

  $S_{married} = [2+,3-]$

  $S_{divorced} = [1+,2-]$

  $Gain(S,marital) = Entropy(S) - (2/10).Entropy(S_{single}) - (5/10).Entropy(S_{married}) - (3/10).Entropy(S_{divorced})$ = 0.97 - (2/10) * 1 - (5/10) * 0.97 - (3/10) * 0.918 = 0.009

- Value (Default) = no, unknown

  $S_{no} = [3+,4-]$

$S_{unknown} = [1+, 2-]$

$Gain(S, Default) = Entropy(S) - (7/10).Entropy(S_{no}) - (3/10).Entropy(S_{unknown})$

= 0.97 - (7/10) * 0.985- (3/10) * 0.528 = 0.1221

- Value (housing) = yes , no

$S_{Yes} = [2+, 4-]$

$S_{No} = [2+, 2-]$

$Gain(S, housing) = Entropy(S) - (6/10).Entropy(S_{Yes}) - (4/10).Entropy(S_{No})$ =

0.97 - (6/10) * 0.918 - (4/10) * 1 = 0.0192

- Value (loan) = Yes ,no

$S_{Yes} = [3+, 0-]$

$S_{No} = [1+, 6-]$

$Gain(S, loan) = Entropy(S) - (3/10).Entropy(S_{Yes}) - (6/10).Entropy(S_{No})]$ = 0.97

- (3/10) * 0 - (7/10) * 0.59 = 0.557

- Value(contact) = cellular, telephone

$S_{cellular} = [3+, 5-]$

$S_{telephone} = [1+, 1-]$

$Gain(S, contact) = Entropy(S) - (8/10).Entropy(S_{cellular}) - (2/10).Entropy(S_{telephone})$

= 0.97 - (8/10) * 0.95 - (2/10) *1 = 0.01

$\Rightarrow$ **Kết luận:**

- Sau khi tính được như vậy ta thấy Gain(S,Loan)=0.557 có giá trị cao nhất nên nó sẽ đứng đầu tiên trong cây.

- Tương tự như vậy ta sẽ tính Gain cho các thuộc tính trong Loan với các feature trong dataset thì ta sẽ nối tiếp được feature nào sẽ nằm bên thuộc tính có giá trị hợp lý so với Loan.

- Cuối cùng ta sẽ thu được một biểu đồ cây và có thể so thuộc tính mới trong dataset để kết luận xem khách hàng cần dự đoán có đăng kí tiền gửi có kì hạn hay không.

Chương 3

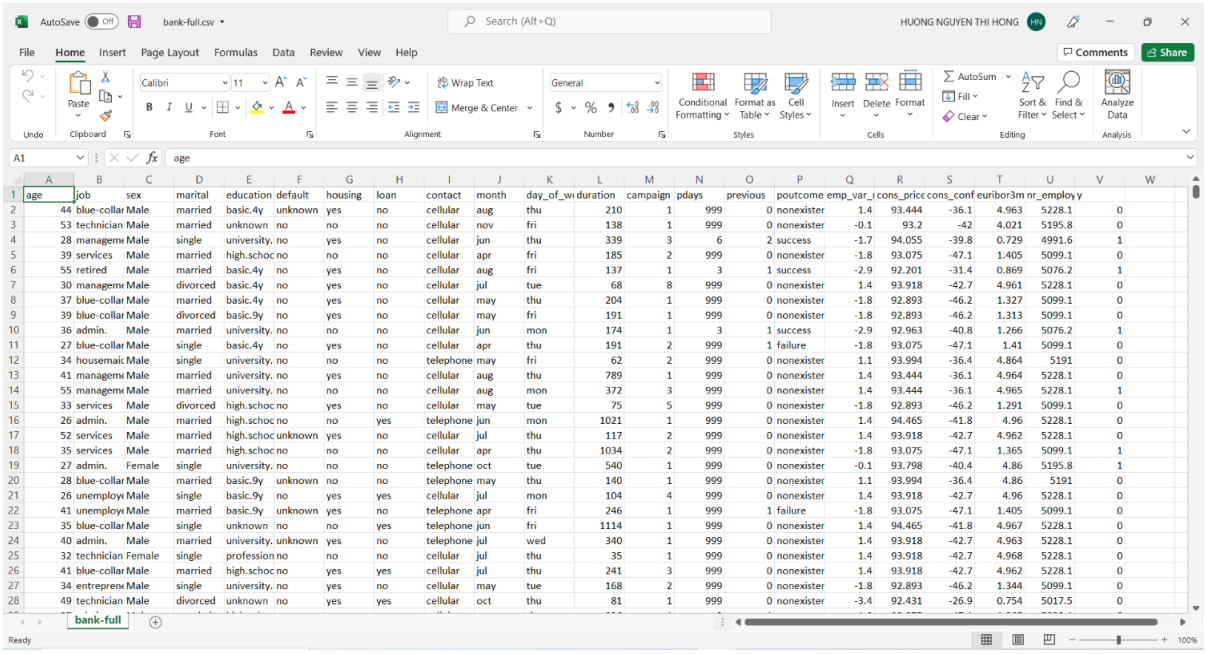
DỮ LIỆU THỰC NGHIỆM

3.1 Giới thiệu tập dữ liệu Bank Marketing

Tập dữ liệu Bank Marketing được xác định và tổng hợp với chiến dịch tiếp thị trực tiếp từ một tổ chức ngân hàng ở Bồ Đào Nha dựa trên các cuộc gọi. Tổ chức này đã cung cấp thông tin được xác thực với các nổ lực tiếp thị thông qua các cuộc gọi điện thoại. Các nổ lực tiếp thị của ngân hàng đều dựa vào khối lượng thông tin khách hàng khổng lồ này để khai thác dữ liệu phù hợp.

Thông thường, cần có nhiều liên hệ với cùng một khách hàng để biết liệu "tiền gửi có kỳ hạn" có được khách hàng đó đăng ký hay không ("yes" hay "no"). Tập dữ liệu được lấy từ trang UCI Machine Learning Repository.

Tập dữ liệu gồm có 41188 trường dữ liệu với 21 thuộc tính đầu vào, cột y là cột chứa thông số "1" (yes) và "0" (no) cho biết khách hàng có đăng ký gửi tiền có kỳ hạn hay không.



age	job	sex	marital	education	default	housing	loan	contact	month	day_of_we	duration	campaign	pdays	previous	poutcome	emp_var_r	cons_price	cons_conf	euribor3m	nr_employ	y
44	blue-collar	Male	married	basic.4y	unknown	yes	no	cellular	aug	thu	210	1	999	0	nonexister	1.4	93.444	-36.1	4.963	5228.1	0
53	technician	Male	married	unknown	no	no	no	cellular	nov	fri	138	1	999	0	nonexister	-0.1	93.2	-42	4.021	5195.8	0
28	managem	Male	single	university.	no	yes	no	cellular	jun	thu	339	3	6	2	success	-1.7	94.055	-39.8	0.729	4991.6	1
39	services	Male	married	high.schoc	no	no	no	cellular	apr	fri	185	2	999	0	nonexister	-1.8	93.075	-47.1	1.405	5099.1	0
55	retired	Male	married	basic.4y	no	yes	no	cellular	aug	fri	137	1	3	1	success	-2.9	92.201	-31.4	0.869	5076.2	1
30	managem	Male	divorced	basic.4y	no	yes	no	cellular	jul	tue	68	8	999	0	nonexister	1.4	93.918	-42.7	4.961	5228.1	0
37	blue-collar	Male	married	basic.4y	no	yes	no	cellular	may	thu	204	1	999	0	nonexister	-1.8	92.893	-46.2	1.327	5099.1	0
39	blue-collar	Male	divorced	basic.9y	no	yes	no	cellular	may	fri	191	1	999	0	nonexister	-1.8	92.893	-46.2	1.313	5099.1	0
36	admin.	Male	married	university.	no	no	no	cellular	jun	mon	174	1	3	1	success	-2.9	92.963	-40.8	1.266	5076.2	1
27	blue-collar	Male	single	basic.4y	no	yes	no	cellular	apr	thu	191	2	999	1	failure	-1.8	93.075	-47.1	1.41	5099.1	0
34	housemaic	Male	single	university.	no	no	no	telephone	may	fri	62	2	999	0	nonexister	1.1	93.994	-36.4	4.864	5191	0
41	managem	Male	married	university.	no	yes	no	cellular	aug	thu	789	1	999	0	nonexister	1.4	93.444	-36.1	4.964	5228.1	0
55	managem	Male	married	university.	no	no	no	cellular	aug	mon	372	3	999	0	nonexister	1.4	93.444	-36.1	4.965	5228.1	1
33	services	Male	divorced	high.schoc	no	yes	no	cellular	may	tue	75	5	999	0	nonexister	-1.8	92.893	-46.2	1.291	5099.1	0
26	admin.	Male	married	high.schoc	no	no	yes	telephone	jun	mon	1021	1	999	0	nonexister	1.4	94.465	-41.8	4.96	5228.1	0
52	services	Male	married	high.schoc	unknown	yes	no	cellular	jul	thu	117	2	999	0	nonexister	1.4	93.918	-42.7	4.962	5228.1	0
35	services	Male	married	high.schoc	no	no	no	cellular	apr	thu	1034	2	999	0	nonexister	-1.8	93.075	-47.1	1.365	5099.1	1
27	admin.	Female	single	university.	no	no	no	telephone	oct	tue	540	1	999	0	nonexister	-0.1	93.798	-40.4	4.86	5195.8	1
28	blue-collar	Male	married	basic.9y	unknown	no	no	telephone	may	thu	140	1	999	0	nonexister	1.1	93.994	-36.4	4.86	5191	0
26	unemploy	Male	single	basic.9y	no	yes	yes	cellular	jul	mon	104	4	999	0	nonexister	1.4	93.918	-42.7	4.96	5228.1	0
41	unemploy	Male	married	basic.9y	unknown	yes	no	telephone	apr	fri	246	1	999	1	failure	-1.8	93.075	-47.1	1.405	5099.1	0
35	blue-collar	Male	single	unknown	no	no	yes	telephone	jun	fri	1114	1	999	0	nonexister	1.4	94.465	-41.8	4.967	5228.1	0
40	admin.	Male	married	university.	unknown	yes	no	telephone	jul	wed	340	1	999	0	nonexister	1.4	93.918	-42.7	4.963	5228.1	0
32	technician	Female	single	profession	no	no	no	cellular	jul	thu	35	1	999	0	nonexister	1.4	93.918	-42.7	4.968	5228.1	0
41	blue-collar	Male	married	high.schoc	no	yes	yes	cellular	jul	thu	241	3	999	0	nonexister	1.4	93.918	-42.7	4.962	5228.1	0
34	entrepren	Male	single	university.	no	yes	no	cellular	may	tue	168	2	999	0	nonexister	-1.8	92.893	-46.2	1.344	5099.1	0
49	technician	Male	divorced	unknown	no	yes	yes	cellular	oct	thu	81	1	999	0	nonexister	-3.4	92.431	-26.9	0.754	5017.5	0

Hình 3.1: Tập dữ liệu *bank-full.csv*

3.2 Đặc tả tập dữ liệu

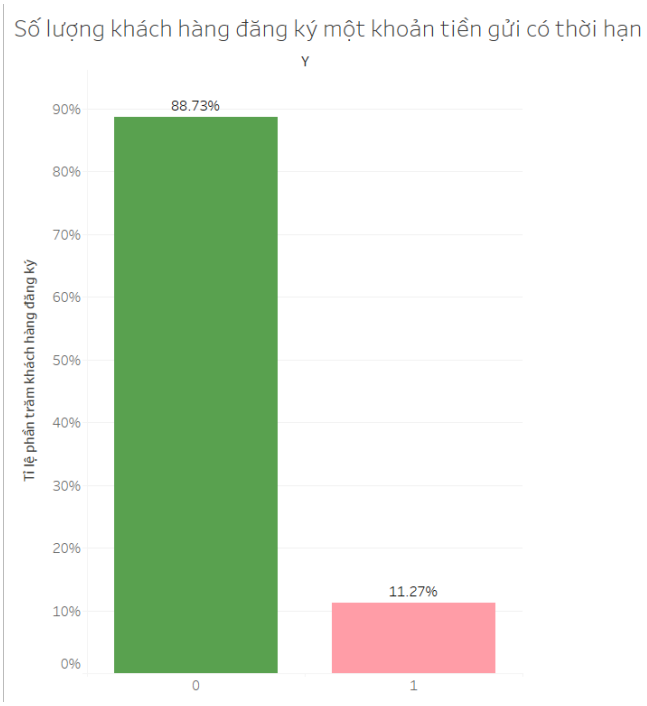
1. age: số tuổi của khách hàng (kiểu dữ liệu số - numeric)
2. job: loại công việc (phân loại:"admin.","blue-collar", "entrepreneur", "housemaid", "management", "retired", "self-employed", "services", "student", "technician", "unemployed", "unknown").
3. sex: giới tính của khách hàng (phân loại: "Male", "Female").
4. marital: tình trạng hôn nhân của khách hàng (phân loại: "divorced", "married", "single", "unknown").
5. education: trình độ giáo dục (phân loại: "basic.4y", "basic.6y", "basic.9y", "high.school", "illiterate", "professional.course", "university.degree", "unknown").
6. default: có tín dụng trong tình trạng vỡ nợ? (phân loại: "yes", "no" và "unknown").
7. housing: có cho vay mua nhà không? (phân loại: "yes", "no" và "unknown").

8. loan: có vay cá nhân không? (phân loại: "yes", "no" và "unknown").
9. contact: kiểu liên lạc (phân loại: "cellular", "telephone").
10. month: tháng liên hệ cuối cùng trong năm (phân loại: “jan”, “feb”, “mar”,.., “nov”, “dec”).
11. day_of_week: ngày liên hệ cuối cùng trong tuần (phân loại: “mon”, “tue”, “wed”, “thu”, “fri”).
12. duration: thời lượng liên hệ cuối cùng, tính bằng giây (số). Lưu ý quan trọng: thuộc tính này ảnh hưởng nhiều đến mục tiêu đầu ra (ví dụ: nếu thời lượng = 0 thì y = ’không’). Khoảng thời gian không được biết trước khi cuộc gọi được thực hiện, cũng như sau khi kết thúc cuộc gọi, y hiển nhiên được biết. Do đó, đầu vào này chỉ nên được đưa vào cho các mục đích chuẩn và nên bị loại bỏ nếu mục đích là có một mô hình dự đoán thực tế.
13. campaign: số liên hệ được thực hiện trong chiến dịch này và cho khách hàng này (kiểu số - numeric, bao gồm liên hệ cuối cùng)
14. pdays: số ngày trôi qua sau khi khách hàng được liên hệ lần cuối từ một chiến dịch trước đó (kiểu dữ liệu số - numeric; 999 nghĩa là khách hàng không có sự tương tác trước đó)
15. previous: số lượng địa chỉ liên hệ của khách hàng được thực hiện trước khi thực hiện việc thu thập dữ liệu (kiểu dữ liệu số - numeric)
16. poutcome: : kết quả của chiến dịch tiếp thị trước đó (phân loại:“failure”, “nonexistent”, “success”")
17. emp.var.rate: tỷ lệ biến động việc làm - chỉ số hàng quý (kiểu dữ liệu số - numeric)
18. cons.price.idx: chỉ số giá tiêu dùng - chỉ số hàng tháng (kiểu dữ liệu số - numeric)
19. cons.conf.idx: chỉ số niềm tin của khách hàng - chỉ số hàng tháng (kiểu dữ liệu - numeric)

20. euribor3m: euribor lãi suất 3 tháng - (kiểu dữ liệu số - numeric)
21. nr.employed: số lượng nhân viên - chỉ số hàng quý (kiểu dữ liệu số - numeric)
22. y - khách hàng đã đăng ký tiền gửi có kỳ hạn chưa? (nhị phân: "0" và "1")

3.3 Trực quan hóa bằng đồ thị

3.3.1 *Số lượng khách hàng đăng ký một khoản tiền gửi có kỳ hạn*



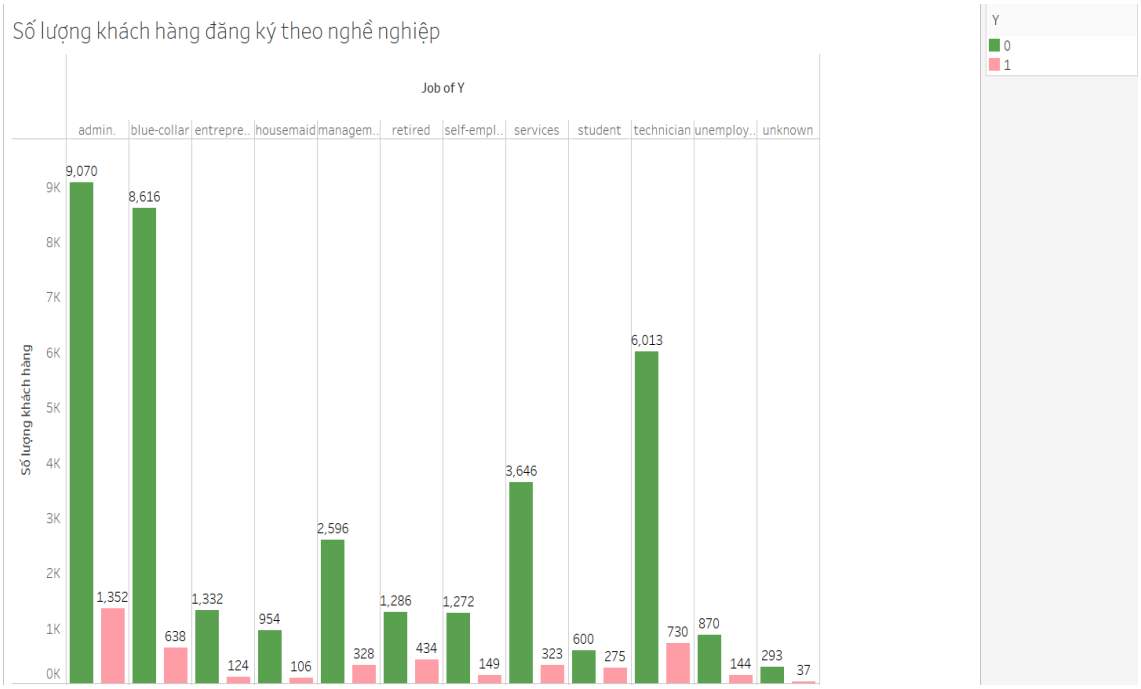
Hình 3.2: Đồ thị biểu diễn tỉ lệ khách hàng đăng ký gửi một khoản tiền có kỳ hạn

⋆ ***Nhận xét***

- Số lượng khách hàng không đăng ký chiếm đa số gần 90 phần trăm, trong khi đó khách hàng có đăng ký chỉ chiếm hơn 11 phần trăm.
- Dữ liệu bị nghiêng về một phía với tỉ lệ 9:1.

- Tập dữ liệu trên không cân bằng (Imbalanced data): mất cân bằng dữ liệu rất phổ biến trong các bài toán phân loại. Việc triển khai một cách cố gắng mô hình phân loại trên dữ liệu không cân bằng như vậy có thể sẽ dẫn đến độ chính xác dự đoán rất thấp. Trong khi đó, tỉ lệ trong bài toán trên lệch hẳn 9:1, cần phải có giải pháp để cải thiện cho mô hình trên bởi vì nếu có đưa ra dự đoán thì kết quả sẽ thiên về phía có phần trăm nhiều hơn dẫn đến dự đoán kém bên phía ít hơn.

3.3.2 *Số lượng khách hàng đăng ký theo nghề nghiệp*



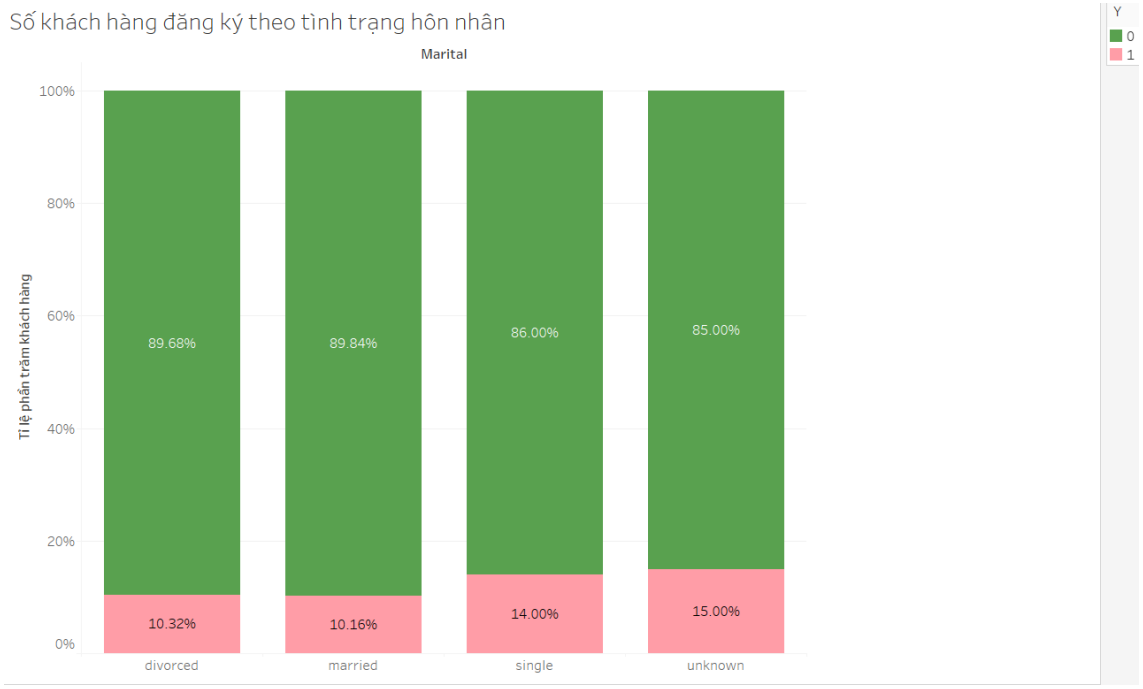
Hình 3.3: Đồ thị biểu diễn số lượng khách hàng đăng ký theo nghề nghiệp

⋆ ***Nhận xét***

- Đồ thị biểu diễn số khách hàng đăng ký thông qua các nghề nghiệp là khác nhau và không đồng đều.
- Admin chiếm giá trị cao nhất trong trường hợp về số người không đăng ký và có đăng ký.

- Unknown chiếm giá trị thấp nhất trong trường hợp về số người đăng ký và không đăng ký.
- 'Job' là một biến độc lập (Independent variable) mang các yếu tố hay điều kiện mà khi thay đổi sẽ ảnh hưởng đến kết quả dự đoán cho y (outcome variable). Đây là biến mang nhiều yếu tố dự đoán cho biến y (kết quả dự đoán của biến y sẽ thay đổi ít nhiều theo biến độc lập).

3.3.3 *Số lượng khách hàng đăng ký theo tình trạng hôn nhân*



Hình 3.4: Đồ thị biểu diễn số phần trăm khách hàng đăng ký theo tình trạng hôn nhân

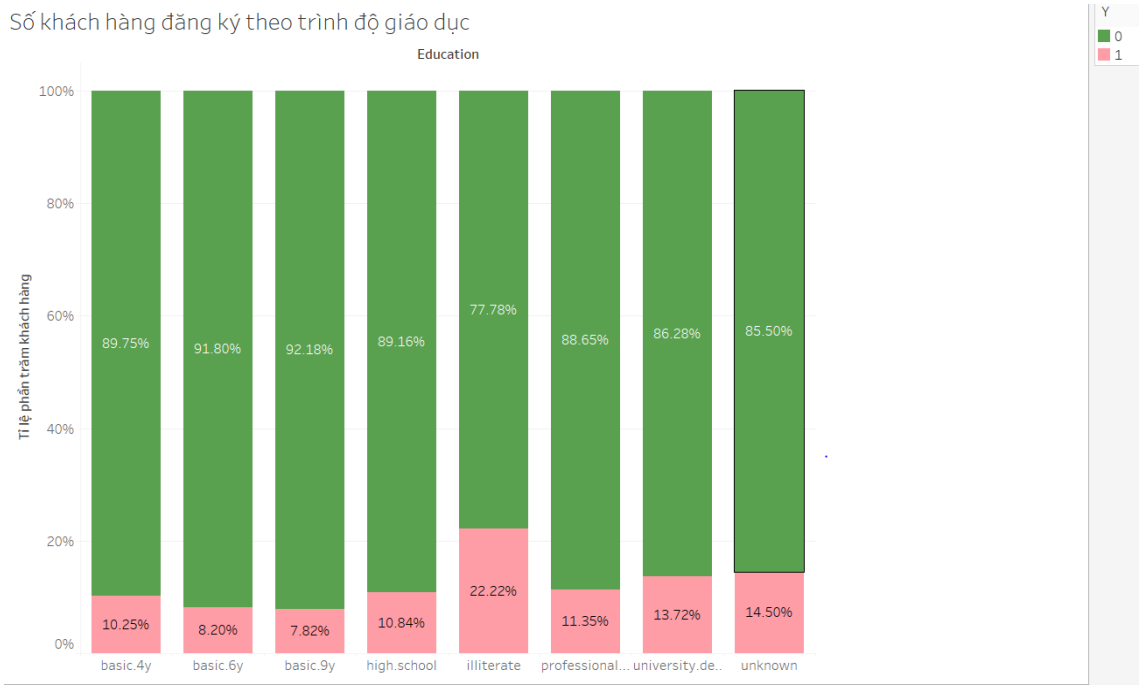
⋆ ***Nhận xét***

- Single và Unknown có tỉ lệ gần như bằng nhau trong trường hợp khách hàng không đăng ký và có đăng ký, chỉ hơn kém 1%.
- Divorced và married: cũng tương tự như trường hợp trên về cả khách hàng đăng ký hay không đăng ký.

- Các tỉ lệ theo tình trạng hôn nhân không chênh lệch nhiều.

⇒ Ta thấy biến độc lập marital không phải là yếu tố dự đoán tốt cho biến outcome y.

3.3.4 *Số lượng khách hàng đăng ký theo trình độ giáo dục*



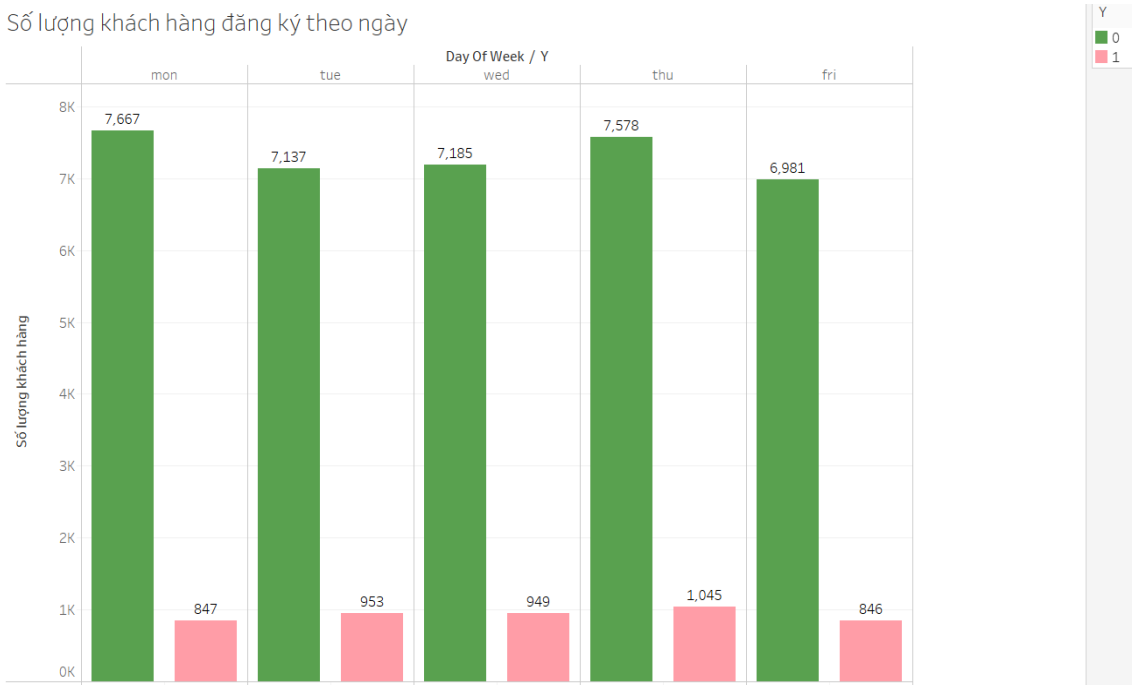
Hình 3.5: Đồ thị biểu diễn số phần trăm khách hàng đăng ký theo trình độ giáo dục

⋆ ***Nhận xét***

- Illiterate có phần trăm số khách hàng đăng ký nhiều nhất và phần trăm khách hàng không đăng ký là thấp nhất.
- Cột basic.4y, basic.6y, basic.9y có số phần trăm khách hàng không đăng ký cao nhất và khách hàng đăng ký thấp nhất.

⇒ Ta thấy biến độc lập education là yếu tố dự đoán tốt cho biến outcome y.

3.3.5 *Số lượng khách hàng đăng ký theo ngày*



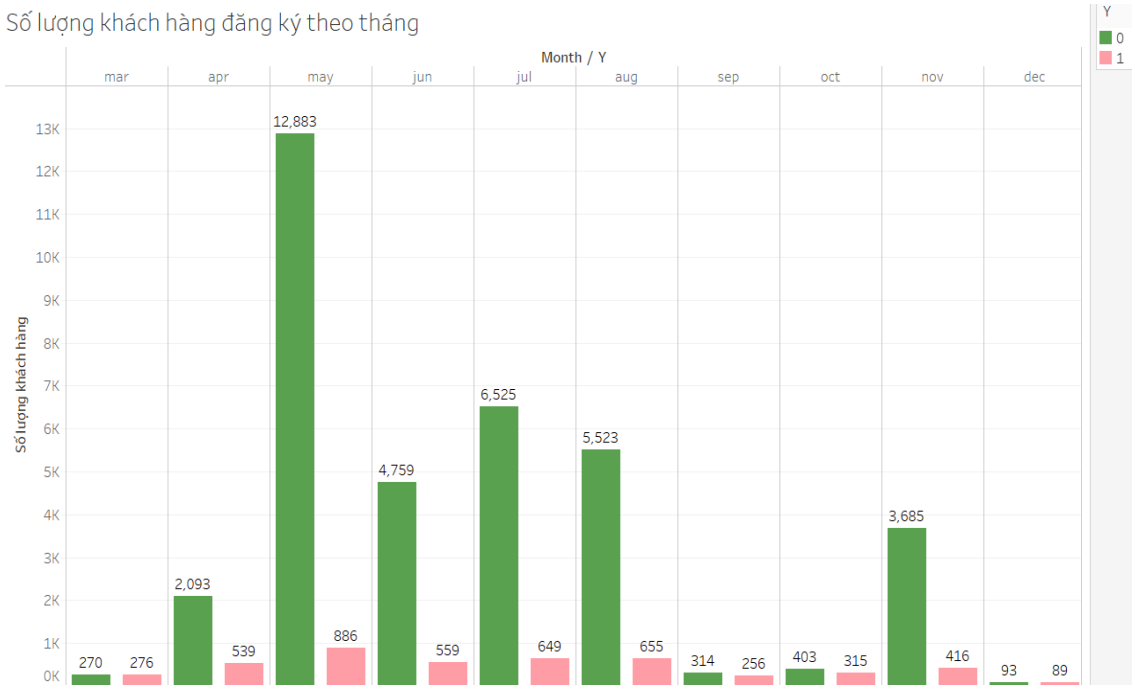
Hình 3.6: Đồ thị biểu diễn số lượng khách hàng đăng ký theo ngày trong tuần

⋆ ***Nhận xét***

- Số lượng khách hàng đăng ký vào các ngày trong tuần cũng không chênh lệch quá nhiều giữa các ngày với nhau.
- Dữ liệu bị nghiêng về phía khách hàng không đăng ký.

⇒ Ta thấy biến độc lập dayofweek không là yếu tố dự đoán tốt cho biến outcome y.

3.3.6 *Số lượng khách hàng đăng ký theo tháng*



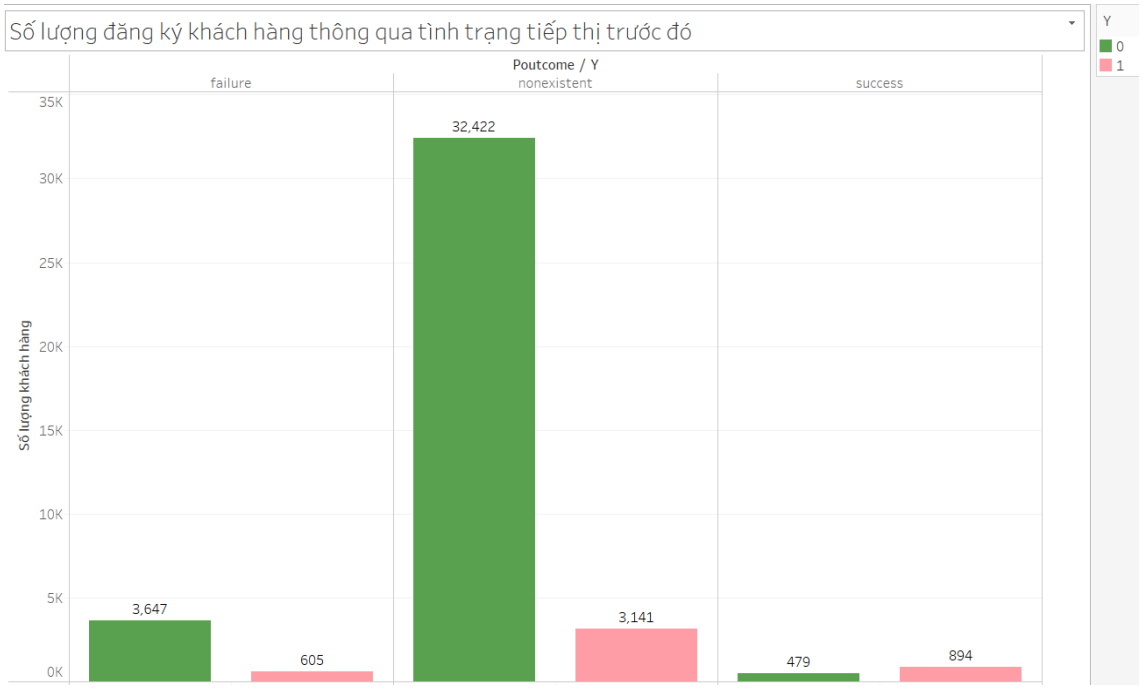
Hình 3.7: Đồ thị biểu diễn số lượng khách hàng đăng ký theo tháng

⋆ ***Nhận xét***

- May có số khách hàng không đăng ký nhiều nhất và số người đăng ký cao nhất.
- Dec thì ngược lại, không đăng ký và có đăng ký đều ít nhất.
- Số lượng biến động khách hàng theo tháng có sự thay đổi không đồng đều.

$\Rightarrow$ Ta thấy biến độc lập month là yếu tố dự đoán tốt cho biến outcome y.

3.3.7 *Số lượng khách hàng đăng ký theo tình trạng tiếp thị trước đó*



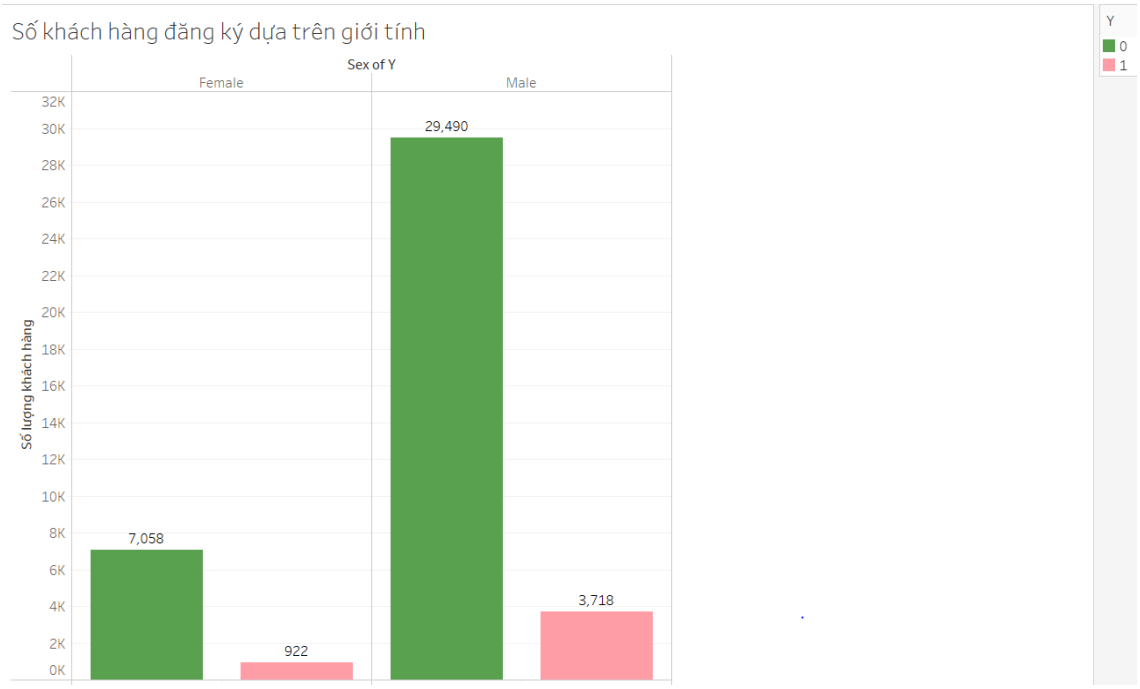
Hình 3.8: Đồ thị biểu diễn số lượng khách hàng đăng ký theo tình trạng tiếp thị trước đó

⋆ ***Nhận xét***

- Nonexistent có số khách hàng không ký và đăng ký chiếm cao nhất.
- Success thì có số khách hàng không đăng ký và đăng ký chiếm thấp nhất.
- Cả 3 tình trạng nhìn chung có sự chênh lệch với nhau.

$\Rightarrow$ Ta thấy biến độc lập poutcome là yếu tố dự đoán tốt cho biến outcome y.

3.3.8 *Số lượng khách hàng đăng ký theo giới tính*



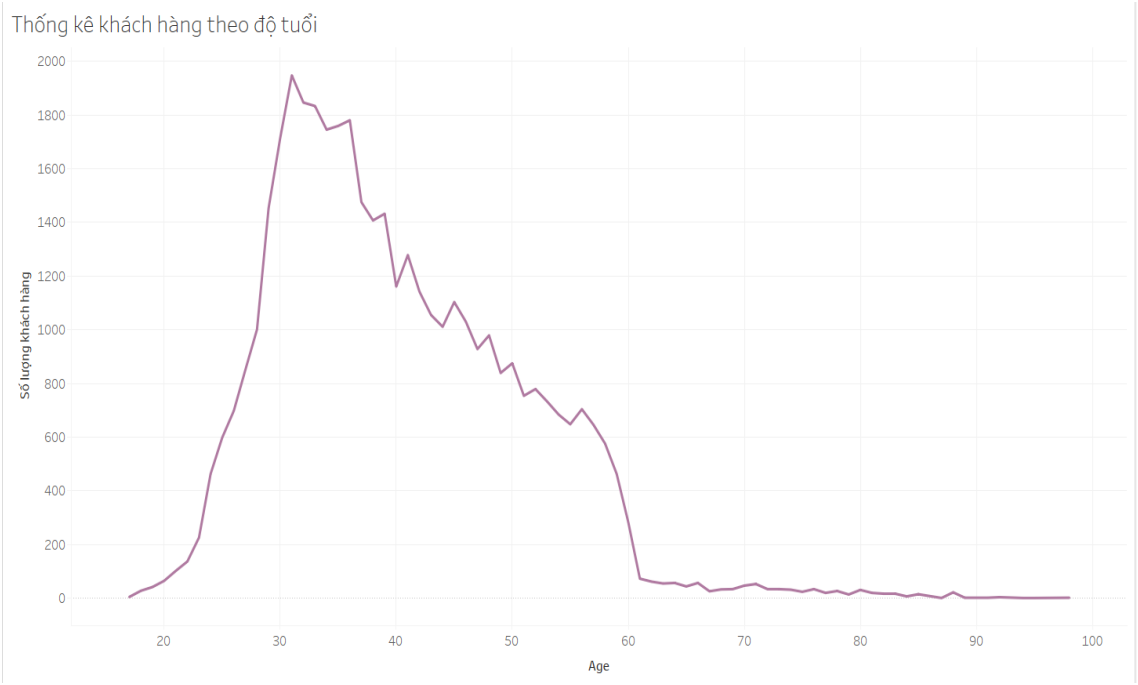
Hình 3.9: Đồ thị biểu diễn số lượng khách hàng đăng ký theo giới tính

⋆ ***Nhận xét***

- Dịch vụ tiếp thị có khách hàng là nam chiếm đa số.
- Giới tính có ảnh hưởng rất nhiều đến chiến dịch tiếp thị trực tiếp này.

$\Rightarrow$ Ta thấy biến độc lập sex là yếu tố dự đoán tốt cho biến outcome y.

3.3.9 *Số lượng khách hàng theo độ tuổi*



Hình 3.10: Đồ thị biểu diễn số lượng khách hàng theo độ tuổi

⋆ ***Nhận xét***

- Hầu hết các khách hàng đến với dịch vụ nằm trong khoảng từ 30 đến 50 tuổi.
- Độ tuổi của khách hàng có ảnh hưởng rất lớn đối với chiến dịch tiếp thị trực tiếp này.

⇒ Ta thấy biến độc lập age là yếu tố dự đoán tốt cho biến outcome y.

KẾT LUẬN:

- Trong tập dữ liệu, các biến độc lập khác nhau mang ảnh hưởng, tác động khác nhau trong việc dự đoán biến y.
- Tập dữ liệu bị mất cân bằng có thể đề ra giải pháp để cân bằng nhằm dự đoán chính xác hơn và cũng nhằm so sánh giữa tập dữ liệu đã cân bằng với tập dữ liệu gốc.

3.4 Xử lý tập dữ liệu

3.4.1 *Quá trình xử lí dữ liệu*

- Đơn giản hóa dữ liệu: Ở biến độc lập *education*, thực hiện việc gôm nhóm các dữ liệu *basic.4y*, *basic.6y* và *basic.9y* thành *basic*.

```
Dữ liệu ban đầu:  ['basic.4y' 'unknown' 'university.degree' 'high.school' 'basic.9y'
 'professional.course' 'basic.6y' 'illiterate']
====================================================
Sau khi xử lý:  ['basic' 'unknown' 'university.degree' 'high.school' 'professional.course'
 'illiterate']
```

Hình 3.11: Dữ liệu trước và sau khi xử lý ở cột education

- Tạo biến giả (Dummy variable):
  - Biến giả là biến độc lập được đưa vào mô hình hồi qui để giải thích các yếu tố định tính.
  - Chuyển đổi biến độc lập loại chuỗi thành các biến độc lập loại số dạng nhị phân (1 hoặc 0).
  - ◇ Ví dụ: Từ biến độc lập marital ta khởi tạo thành một biến giả dạng nhị phân
    marital = married
    marital-married = 1 nếu khách hàng đó có kết hôn
    marital-married = 0 nếu khách hàng không có kết hôn
- Các biến độc lập cần xử lý dữ liệu:
  - Sex:
    “Male”, “Female"
  - Job:
    “admin”, “blue-collar”, “entrepreneur”, “housemaid”,
    “management”, “retired”, “self-employed”, “services”,
    “student”, “technician”, “unemployed”, “unknown”
  - marital:
    “divorced”, “married”, “single”, “unknown”

- education:

  “basic”, “high.school”, “illiterate”, “professional.course”, “university.degree”, “unknown”

- default:

  “no”, “yes”, “unknown”

- housing:

  “no”, “yes”, “unknown”

- loan:

  “no”, “yes”, “unknown”

- contact:

  “cellular”, “telephone”

- month:

  “jan”, “feb”, “mar”,..., “nov”, “dec”

- day_of_week:

  “mon”, “tue”, “wed”, “thu”, “fri”

- poutcome:

  “failure”, “nonexistent”, “success”

```
   age          job   sex  marital  ... cons_conf_idx euribor3m nr_employed
0   44  blue-collar  Male  married  ...         -36.1     4.963      5228.1
1   53   technician  Male  married  ...         -42.0     4.021      5195.8
2   28   management  Male   single  ...         -39.8     0.729      4991.6
3   39     services  Male  married  ...         -47.1     1.405      5099.1
4   55      retired  Male  married  ...         -31.4     0.869      5076.2

[5 rows x 22 columns]
['age' 'job' 'sex' 'marital' 'education' 'default' 'housing' 'loan'
 'contact' 'month' 'day_of_week' 'duration' 'campaign' 'pdays' 'previous'
 'poutcome' 'emp_var_rate' 'cons_price_idx' 'cons_conf_idx' 'euribor3m'
 'nr_employed' 'y']
```

Hình 3.12: Dữ liệu trước khi tạo biến giả

```
   age  duration  ...  poutcome_nonexistent  poutcome_success
0   44       210  ...                     1                 0
1   53       138  ...                     1                 0
2   28       339  ...                     0                 1
3   39       185  ...                     1                 0
4   55       137  ...                     0                 1

[5 rows x 64 columns]
['age' 'duration' 'campaign' 'pdays' 'previous' 'emp_var_rate'
 'cons_price_idx' 'cons_conf_idx' 'euribor3m' 'nr_employed' 'y'
 'job_admin.' 'job_blue-collar' 'job_entrepreneur' 'job_housemaid'
 'job_management' 'job_retired' 'job_self-employed' 'job_services'
 'job_student' 'job_technician' 'job_unemployed' 'job_unknown'
 'sex_Female' 'sex_Male' 'marital_divorced' 'marital_married'
 'marital_single' 'marital_unknown' 'education_basic'
 'education_high.school' 'education_illiterate'
 'education_professional.course' 'education_university.degree'
 'education_unknown' 'default_no' 'default_unknown' 'default_yes'
 'housing_no' 'housing_unknown' 'housing_yes' 'loan_no' 'loan_unknown'
 'loan_yes' 'contact_cellular' 'contact_telephone' 'month_apr' 'month_aug'
 'month_dec' 'month_jul' 'month_jun' 'month_mar' 'month_may' 'month_nov'
 'month_oct' 'month_sep' 'day_of_week_fri' 'day_of_week_mon'
 'day_of_week_thu' 'day_of_week_tue' 'day_of_week_wed' 'poutcome_failure'
 'poutcome_nonexistent' 'poutcome_success']
kích thước: 64
```

Hình 3.13: Dữ liệu sau khi tạo biến giả

3.4.2 *Phương pháp SMOTE*

- Tạo dữ liệu mẫu giả cho tập thiểu số sao cho số phần tử của nó được nhiều lên bằng cách là lặp lại mỗi điểm trong nhóm thiểu số nhiều lần.
- SMOTE (Synthetic Minority Over-sampling): là phương pháp sinh mẫu nhằm gia tăng kích thước mẫu của nhóm thiểu số trong trường hợp xảy ra mất cân bằng mẫu. Để gia tăng kích thước mẫu, với mỗi một mẫu thuộc nhóm thiểu số sẽ lựa chọn ra k mẫu láng giềng gần nhất với nó và sau đó thực hiện tổ hợp tuyến tính để tạo ra mẫu giả lập.

```
------------Oversampling - SMOTE------------
Số lượng tập train: 28831
Số lượng tập test: 12357
length Oversampling: 51134
Số lượng khách hàng không đăng ký trong tập Oversampling: 25567
Số lượng khách hàng đăng ký trong tập Oversampling: 25567
Tỉ lệ khách hàng không đăng ký trong tập Oversampling: 0.5
Tỉ lệ khách hàng đăng ký trong tập Oversampling: 0.5
```

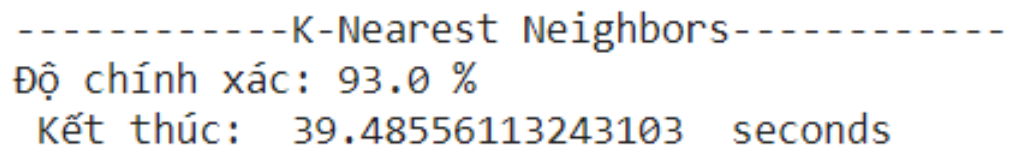
Hình 3.14: Dữ liệu sau khi dùng SMOTE

Chương 4

THỰC NGHIỆM

4.1 Thuật toán KNN

- Tập dữ liệu gốc:



Hình 4.1: Kết quả sau khi áp dụng thuật toán

```
_____________________________________________
Vị trí dữ liệu dự đoán:  4572
Dự đoán:  1.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  17662
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  18632
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  6439
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  11000
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  11172
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  10213
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  92
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  17925
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  12868
Dự đoán:  0.0
Kết quả đúng:  0.0
Độ chính xác 10 tập test ngẫu nhiên:  90.0 %
```

Hình 4.2: Kết quả dự đoán 10 tập ngẫu nhiên

- Tập dữ liệu đã được cân bằng: chia dữ liệu thành hai tập train, tỉ lệ 7:3.

```
------------K-Nearest Neighbors------------
Độ chính xác: 89.0 %
 Kết thúc:  40.66314125061035  seconds
```

Hình 4.3: Kết quả sau khi áp dụng thuật toán với dữ liệu cân bằng

```
_____________________________________________
Vị trí dữ liệu dự đoán:  5661
Dự đoán:  1.0
Kết quả đúng:  1.0
_____________________________________________
Vị trí dữ liệu dự đoán:  20819
Dự đoán:  1.0
Kết quả đúng:  1.0
_____________________________________________
Vị trí dữ liệu dự đoán:  20274
Dự đoán:  1.0
Kết quả đúng:  1.0
_____________________________________________
Vị trí dữ liệu dự đoán:  4622
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  14461
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  9723
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  15013
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  15992
Dự đoán:  1.0
Kết quả đúng:  1.0
_____________________________________________
Vị trí dữ liệu dự đoán:  17982
Dự đoán:  0.0
Kết quả đúng:  0.0
_____________________________________________
Vị trí dữ liệu dự đoán:  8185
Dự đoán:  1.0
Kết quả đúng:  1.0
Độ chính xác 10 tập test ngẫu nhiên:  100.0 %
```

Hình 4.4: Kết quả dự đoán 10 tập ngẫu nhiên

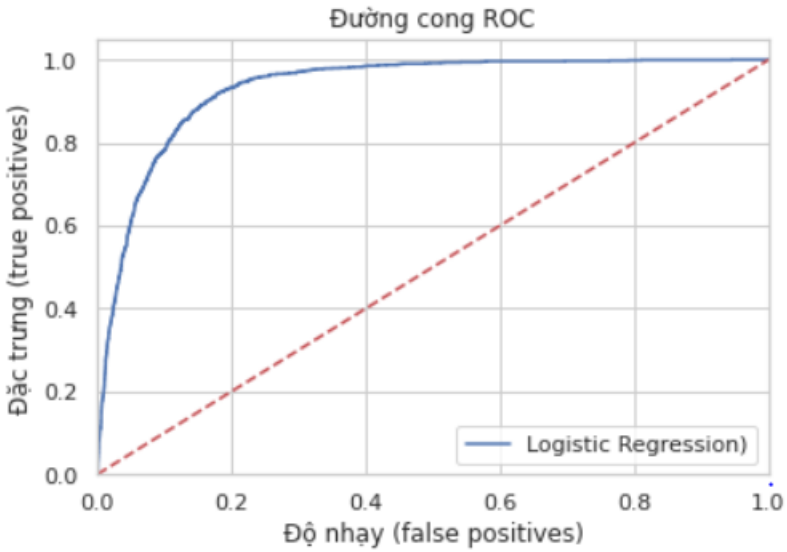
- Nhận xét:
  - Dữ liệu gốc khi tiến hành thực thi thuật toán lại có độ chính xác cao hơn đối với tập cân bằng với phương pháp SMOTE.
  - Khi chọn K ta cần phải lưu ý vào tập dữ liệu và số lượng dữ liệu vì nếu chọn K quá nhỏ thì dữ liệu dự đoán sẽ bị nhiều.
  - Kết quả dự đoán 10 tập ngẫu nhiên ở dữ liệu đã cân bằng lại mang chính xác 100% với dữ liệu gốc chỉ 90%

4.2 Thuật toán hồi quy logistic

- Tập dữ liệu gốc:

```
------------Logistic Regression------------
Số lượng tập train: 28831
Số lượng tập test: 12357
Độ chính xác:  91.0 %
 Kết thúc:  8.422946691513062  seconds
============================================================
```

Hình 4.5: Kết quả sau khi áp dụng thuật toán

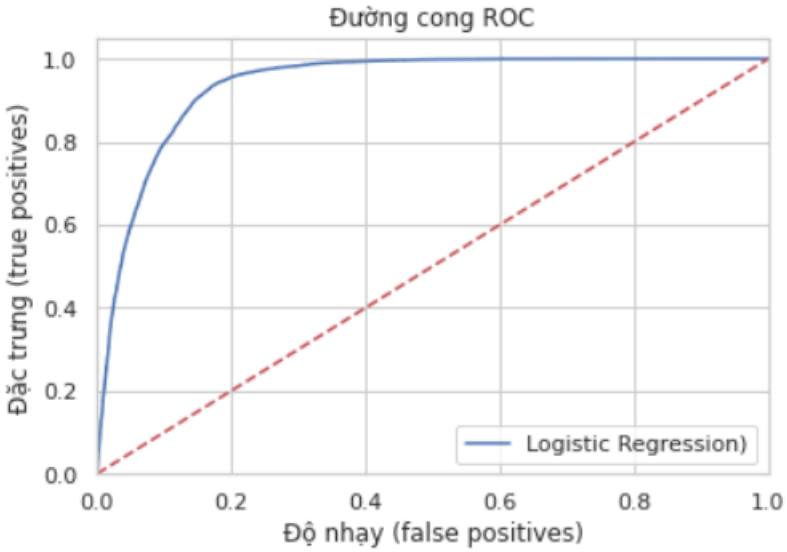


Hình 4.6: Đồ thị đường cong ROC

- Tập dữ liệu đã được cân bằng: chia dữ liệu thành hai tập train, tỉ lệ 7:3.

```
Độ chính xác:  91.0 %
 Kết thúc:  38.58591938018799  seconds
============================================================
```

Hình 4.7: Kết quả sau khi áp dụng thuật toán với dữ liệu cân bằng



Hình 4.8: Đồ thị đường cong ROC với dữ liệu cân bằng

- Nhận xét:
  - Dữ liệu khi tiến hành thực thi thuật toán ta thấy dữ liệu cân bằng với dữ liệu gốc có độ chính xác bằng nhau.
  - Đường cong ROC dùng để đánh giá kết quả của một dự đoán, ta thấy đường cong của cả hai trường hợp đều có dấu hiệu đi lên dọc theo bìa trái chứng tỏ thuật toán áp dụng có độ chính xác rất tốt.

4.3 Thuật toán Naive Bayes

- Tập dữ liệu gốc:

```
------------Naive Bayes------------
Độ chính xác:  86.0 %
 Kết thúc:  0.04590344429016113  seconds
```

Hình 4.9: Kết quả sau khi áp dụng thuật toán

- Tập dữ liệu đã được cân bằng: chia dữ liệu thành hai tập train, tỉ lệ 7:3.

```
==========================================================
------------Naive Bayes------------
Độ chính xác:  85.0 %
 Kết thúc:  0.06400942802429199  seconds
```

Hình 4.10: Kết quả sau khi áp dụng thuật toán với dữ liệu cân bằng

- Nhận xét:
  - Dữ liệu khi tiến hành thực thi thuật toán ta thấy dữ liệu cân bằng với dữ liệu gốc có độ chính xác gần bằng nhau, hơn nhau chỉ 1 %
  - Độ chính xác của thuật toán cũng khá cao nhưng không cao bằng hai thuật toán trước đó.

4.4 Thuật toán Support Vector Machine

- Tập dữ liệu gốc:

```
------------SVM------------
Độ chính xác:  90.0 %
 Kết thúc:  112.4797830581665  seconds
```

Hình 4.11: Kết quả sau khi áp dụng thuật toán

- Tập dữ liệu đã được cân bằng: chia dữ liệu thành hai tập train, tỉ lệ 7:3.

```
------------SVM------------
Độ chính xác:  90.0 %
 Kết thúc:  279.8825612068176  seconds
```

Hình 4.12: Kết quả sau khi áp dụng thuật toán với dữ liệu cân bằng

- Nhận xét:
  - Dữ liệu khi tiến hành thực thi thuật toán ta thấy dữ liệu cân bằng với dữ liệu gốc có độ chính xác giống nhau.

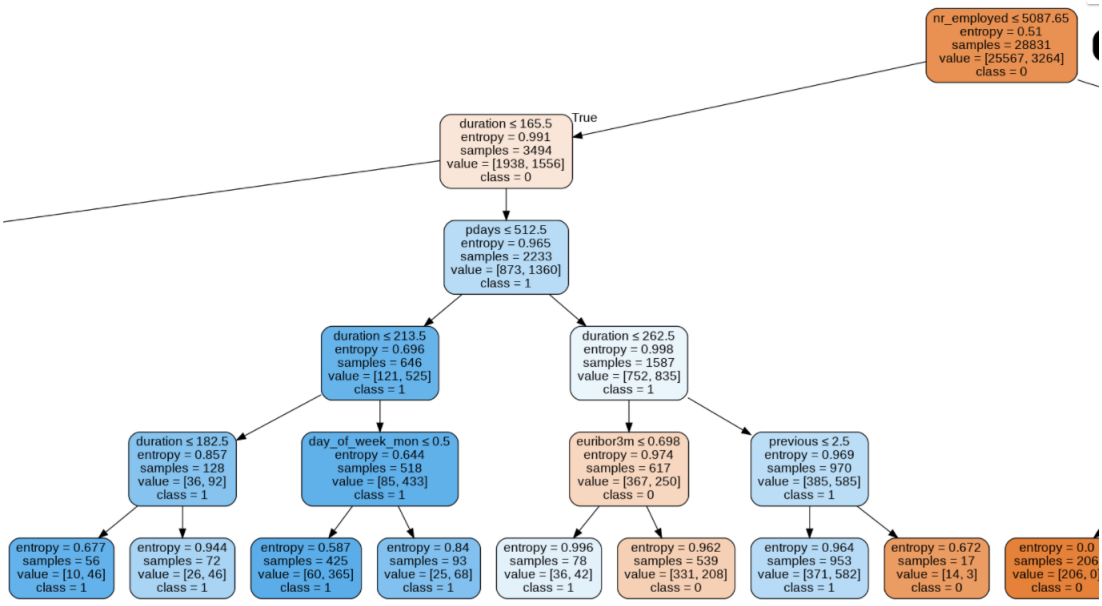
- Độ chính xác của thuật toán cao, tuy nhiên thời gian thực thi lại chậm hơn rất nhiều so với các thuật toán trước đó.

4.5 Thuật toán Decision Tree

- Tập dữ liệu gốc:

```
------------Decision tree------------
Độ chính xác:  92.0 %
 Kết thúc:  0.10997319221496582  seconds
```

Hình 4.13: Kết quả sau khi áp dụng thuật toán



Hình 4.14: Hình ảnh một phần của cây quyết định với tập dữ liệu gốc

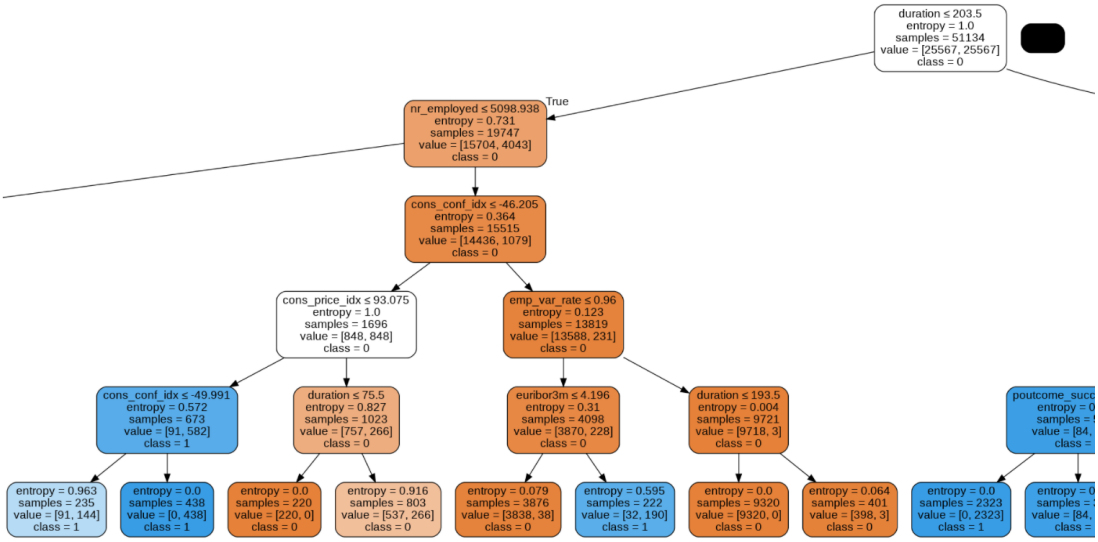
- Chọn thuộc tính nr_employed làm gốc vì nó có giá trị Gain lớn nhất.
- samples ở đây chứa 28831 có nghĩa là cây này được training 28831 mẫu dữ liệu.
- Value [25567, 3264] là giá trị lớp mẫu, tức là có 25567 thuộc lớp 0, còn lại phân vào lớp 1. Lớp 0 có giá trị lớn hơn nên gốc cây này phân vào lớp 0.

- So sánh điều kiện với nr_employed nếu đúng thì nó sẽ phân nhánh ra cây con mới bên trái.

- Tập dữ liệu đã được cân bằng: chia dữ liệu thành hai tập train, tỉ lệ 7:3.

```
------------Decision tree------------
Độ chính xác:  86.0 %
 Kết thúc:  0.26703691482543945  seconds
```

Hình 4.15: Kết quả sau khi áp dụng thuật toán với dữ liệu cân bằng



Hình 4.16: Hình ảnh một phần của cây quyết định với tập dữ liệu cân bằng

- Chọn thuộc tính duration làm gốc vì nó có giá trị Gain lớn nhất.
- samples ở đây chứa 51134 có nghĩa là cây này được training 28831 mẫu dữ liệu.
- Value [25567, 25567] là giá trị lớp mẫu, tức là có 25567 thuộc lớp 0, còn lại phân vào lớp 1. Hai giá trị bằng nhau phân gốc này thành lớp 0.
- So sánh điều kiện với duration nếu đúng thì nó sẽ phân nhánh ra cây con mới bên trái.

- Nhận xét:

- Dữ liệu khi tiến hành thực thi thuật toán ta thấy dữ liệu cân bằng với dữ liệu gốc có độ chính xác khá chênh lệch.
- Độ chính xác của thuật toán với tập dữ liệu gốc cao hơn so với tập dữ liệu được cân bằng.
- Tính đến thời điểm này, thì thuật toán Decision Tree thực nghiệm trên tập dữ liệu gốc có độ chính xác cao với thời gian chạy nhanh nhất so các thuật toán trước.

⇒ **Kết luận:**

- Đối với tập dữ liệu gốc, sau khi trải qua 5 thuật toán thì ta thấy Decision Tree có thời gian chạy nhanh nhất và độ chính xác thấp hơn KNN 1% nhưng thời gian chạy KNN lại lâu hơn.
- Dối với tập dữ liệu được xử lý cân bằng, thuật toán LR có độ chính xác cao nhất. Thuật toán NB với độ chính xác 85% cho ra kết quả nhanh nhất.

Chương 5

KẾT LUẬN

5.1 Kết quả đạt được

- Thu thập và xử lý tập dữ liệu phù hợp cho việc nghiên cứu báo cáo.
- Thực hiện dự đoán với tập dữ liệu ngân hàng của dịch vụ khách hàng đăng ký tiền gửi có kỳ hạn.
- Đồ thị hóa với Tableau.
- Tìm hiểu 5 thuật toán và áp dụng các thuật toán ấy vào đề tài để đưa ra kết quả dự đoán.

5.2 Mặt hạn chế

Trong quá trình làm báo cáo, chúng em đã cố gắng nghiên cứu và tìm hiểu để có được những kết quả trên. Tuy nhiên do hạn chế về mặt kiến thức cũng như kĩ năng,mặt khác về một số điều kiện về thời gian và không gian, chúng em chưa thể sử dụng nhiều thuật toán khác để áp dụng.

5.3 Hướng phát triển

Vì đề tài còn mang tính thực tế hơn nữa, có khả năng ứng dụng cao với các mục tiêu cho tương lai sau này như dùng các thuật toán học máy khác nhau để phân tích và dự đoán tìm ra các chiến thuật hiệu quả nhất trong chiến dịch tiếp thị ngân hàng ví dụ như dự đoán hành vi trả nợ đúng hạn của khách hàng, đăng ký dịch vụ thẻ tín dụng,... để đưa ra các chiến lược hiệu quả và mang đến lợi nhuận cao cho ngân hàng.

Chương 6

TÀI LIỆU THAM KHẢO

1. Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.
2. Draper, N. R., & Smith, H. (1998). "Dummy" variables. Applied regression analysis, 299-325.
3. Jason Brownlee. (2020). SMOTE for Imbalanced Classification with Python. Machine Learning Mastery.
4. Hadeer Hammad & Min Zhou. (2020). Dữ liệu không cân bằng trong phân loại: Giải pháp chung & Nghiên cứu điển hình. ICHI.PRO.
5. huytuong010101. (2020). Thuật Toán K-Nearest Neighbors (KNN). codelearn.io.
6. HVIT CLAN. (2020). Bài 8 - KNN (K-Nearest Neighbors). hvitclan.vn.
7. Kiến thức kinh tế. (2020). Thuật toán K láng giềng gần nhất (K-Nearest Neighbor - KNN) là gì? vietnambiz.vn.
8. BIGDATAUNI. (2021). Tổng quan về Logistic Regresstion - Phần 2. bigdatauni.com
9. BIGDATAUNI. (2021). Tổng quan về Logistic Regresstion - Phần 1. bigdatauni.com.
10. Blog Trí tuệ nhân tạo. (2019). Cây Quyết Định (Decision Tree). trituenhantao.io.

11. ICHI.PRO. (2020). 5 kỹ thuật SMOTE để lấy mẫu quá mức cho dữ liệu mất cân bằng của bạn. ichi.pro.

12. Data Science Notebook. (2016). Kernel Methods & Non Linear SVM. Binary Classification.

13. VIBLO. (2020). Giới thiệu về Support Vector Machine (SVM). viblo.asia.

14. Data Science, Machine Learning. (2018). Naive Bayes Classification (NBC) là gì?. 1upnote.me.

15. vietnambiz. (2019). Biến giả (Dummy variable) trong phân tích hồi qui là gì?. vietnambiz.vn.