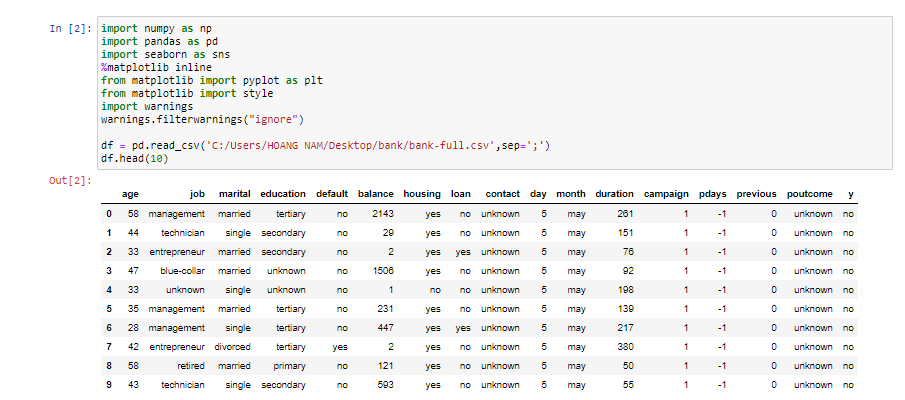
Data Exploration

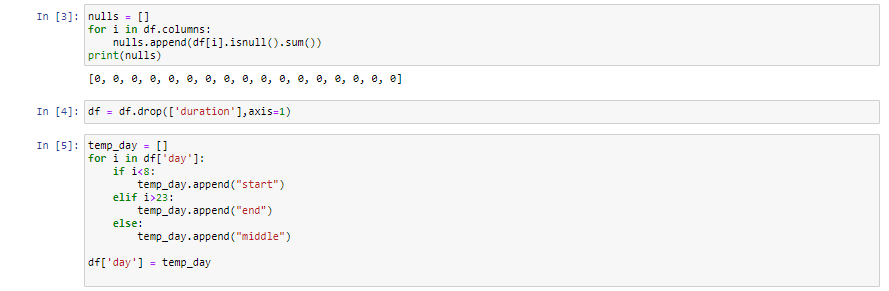
Đọc dữ liệu từ file excel



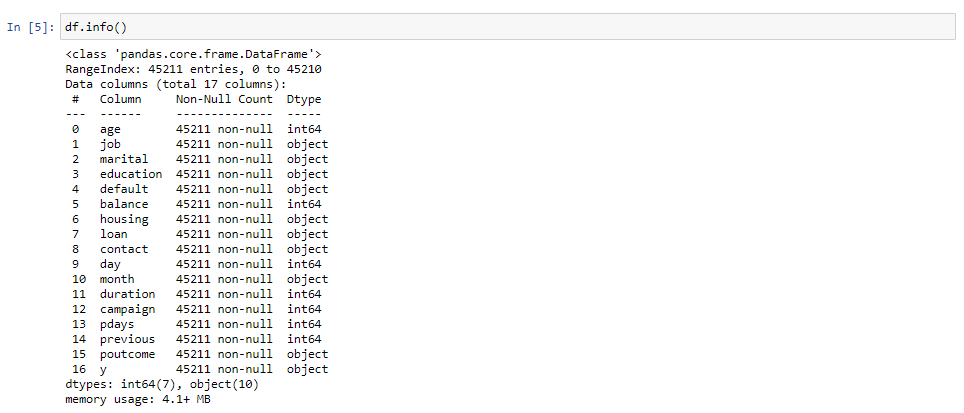
Kiểm tra file data có bao nhiêu dòng và cột



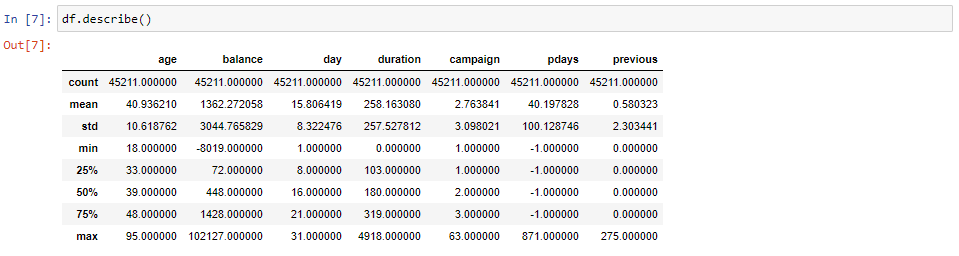
Kiểm tra số dòng dữ liệu rỗng của từng cột



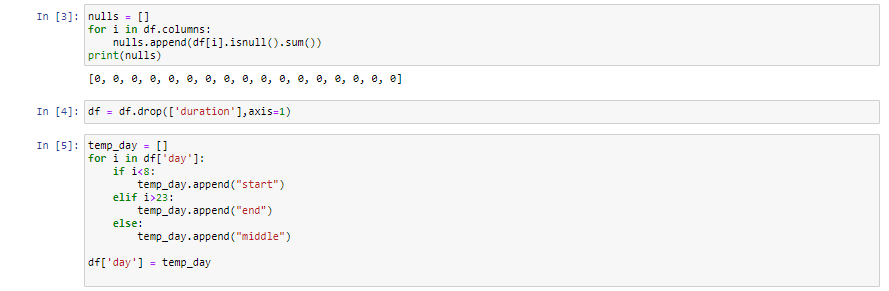
Kiểm tra kiểu dữ liệu của từng biến



Coi nhưng giá trị mô tả cơ bản của từng cột

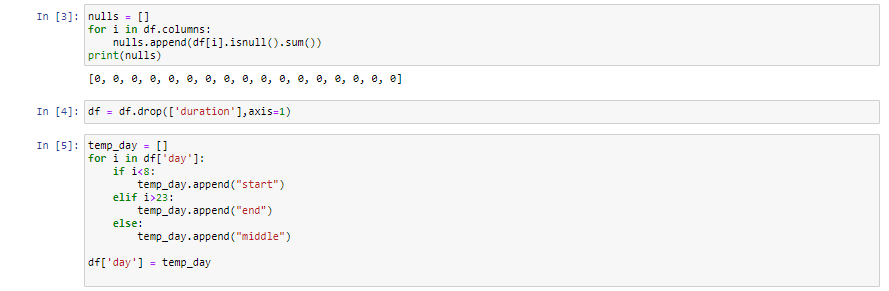


Trước khi tiếp tục, chúng tôi sẽ xóa cột "duration" khỏi tập dữ liệu của mình vì đây là thông tin mà lẽ ra chúng ta chưa có. Trường này ảnh hưởng nhiều đến kết quả đầu ra vì nếu một người có thời lượng cuộc gọi bằng 0, chúng ta có thể cho rằng người gọi không nhấc máy và do đó không có cơ hội bán được hàng. Ngược lại, nếu thời lượng cuộc gọi lớn đồng nghĩa khách hàng thật sự rất quan tâm đến sản phẩm và cơ hội bán được sẽ gần như 100%.



# **Data Cleaning: Numeric Variables**

Cột ‘day’ thể hiện theo ngày trong tuần, chúng ta thực sự nên coi đây là một biến số chứ không phải một số. Ngoài ra, chúng ta sẽ kết hợp các giá trị này thành 3 nhóm: đầu tháng, giữa tháng và cuối tháng. Điều này sẽ ngăn chúng ta có 31 biến được mã hóa khác nhau và tổng quát hóa biến để nó có ý nghĩa đối với vấn đề kinh doanh hiện tại. Sau khi chuyển đổi "day" thành biến số, chúng tôi sẽ mã hóa một biến này bằng các biến phân loại khác .



Biến "pdays" đại diện cho số ngày đã trôi qua sau khi khách hàng được liên hệ lần cuối từ một chiến dịch trước đó. Chúng tôi cũng cần lưu ý rằng nếu khách hàng chưa từng được liên hệ, giá trị của trường này là -1. Trở thành giá trị -1 không có ý nghĩa trong trường này và sẽ không nắm bắt chính xác rằng khách hàng không được liên hệ Tôi thường sẽ tạo một biến nhị phân mới nhưng "poutcome" = nonexistant nắm bắt điều này cho chúng tôi.

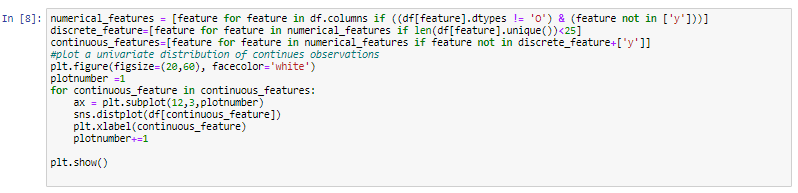
Như chúng ta đã thấy từ các thống kê ở trên về các biến số, có những danh sách tiềm năng. Hãy nói về một số giá trị tối đa cho một số biến của chúng tôi:

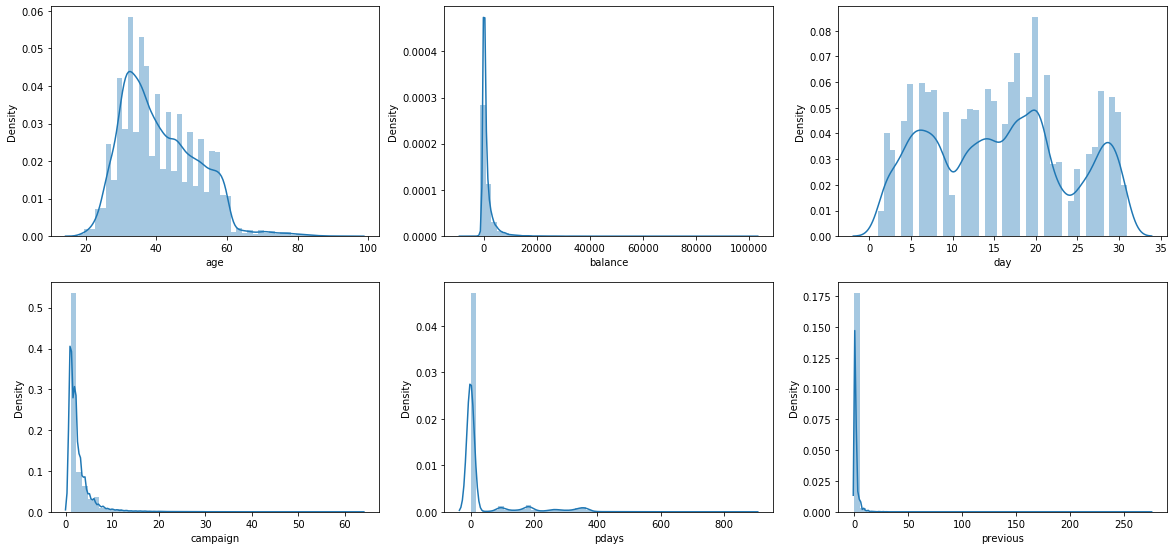
1 - **Balance**: Việc có 81.000 đô la trong tài khoản séc của bạn không hẳn là chưa từng thấy. Ví dụ, điều này sẽ có ý nghĩa nếu một người có mức lương cao, xen kẽ với việc tiết kiệm.

2 - **pdays**: Hơi bất thường khi liên hệ với khách hàng tiềm năng sau 854 ngày (gần 2,5 năm) mặc dù điều đó là hợp lý. Ví dụ: có thể sau 100 ngày, họ đã bị xóa khỏi danh sách tiếp thị nhưng một lá cờ đã được kéo lên có thể cho biết rằng người đó thích thú sẽ mở tài khoản và do đó được đưa trở lại danh sách tiếp thị. Mặc dù chúng đã bị xóa sau đó được thêm lại, nhưng có thể chúng không ngừng đếm số ngày ở giữa.

3 - **previous**: Giống như các biến trước mà chúng ta đã thảo luận, biến này cũng hợp lý. Nếu một khách hàng ở trong nhiều chiến dịch kéo dài trong nhiều quý hoặc thậm chí nhiều năm thì chắc chắn có thể thêm tới hơn 50 địa chỉ liên hệ.

4 -  **Campaign**: Việc tranh giành hơn 60 lần cho một chiến dịch tiếp thị là điều hơi chưa từng xảy ra ngay cả đối với một chiến dịch kéo dài một năm. Hãy thử nghĩ xem, sau khi tiếp cận với một người 20 lần, bạn có nghĩ rằng họ có thể không quan tâm? Ngoài ra, hầu hết các công ty sẽ không lãng phí tiền như vậy. Điều này sẽ cần điều tra thêm.



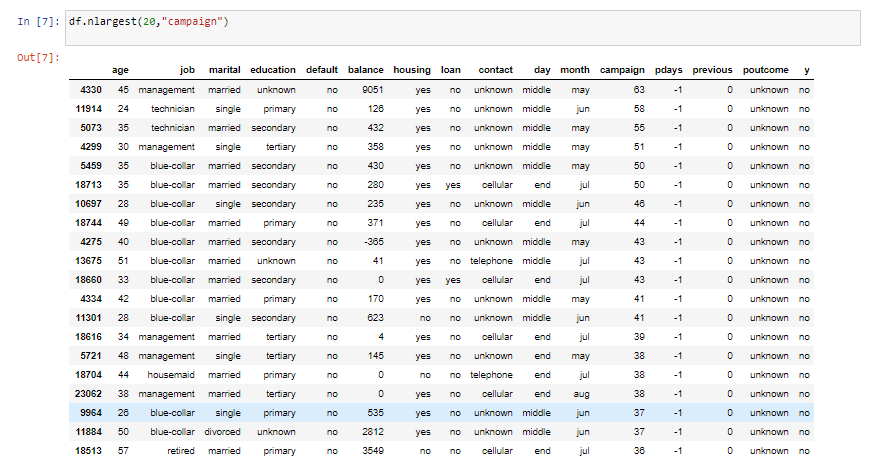


Ta thấy, có vẻ như age, day được phân phối bình thường

Balance, compaign, pdays and previous  bị lệch về bên trái và dường như có một số ngoại lệ.

# **Outliar Analysis**

Sắp xếp giảm “**campaign”**



Dựa trên dữ liệu trên, không có dấu hiệu cho thấy có sai sót trong dữ liệu trên. Có vẻ như tất cả các dữ liệu trên đều có mà trước đây chúng không được liên hệ với nhau. Tôi có thể cho rằng công ty muốn thực hiện một cách tiếp cận linh hoạt để tiếp thị cho các khách hàng tiềm năng mới. Vì lý do này, tôi sẽ không đưa ra bất kỳ quan sát nào mặc dù tôi sẽ đảm bảo ghi nhớ điều này. Mặc dù, tôi quyết định không loại bỏ bất kỳ biến nào, nhưng chắc chắn có một trường hợp mạnh mẽ được thực hiện để chuyển các hàng hoặc chuyển đổi các cột để điều chỉnh cho các giá trị ngoại lệ.

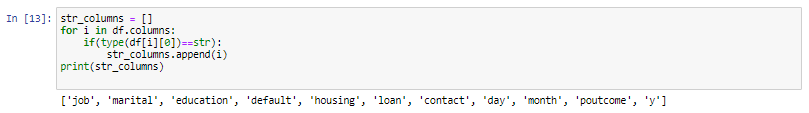
# **Data Cleaning: Catigorical Variables**

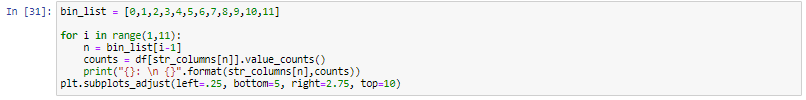
Chọn những cột có kiểu dữ liệu String



## Adjusting the Binary Categorical Variables

## Lấy ra các biến số kiểu string



Số liệu  


job:

blue-collar 9732

management 9458

technician 7597

admin. 5171

services 4154

retired 2264

self-employed 1579

entrepreneur 1487

unemployed 1303

housemaid 1240

student 938

unknown 288

Name: job, dtype: int64

marital:

married 27214

single 12790

divorced 5207

Name: marital, dtype: int64

education:

secondary 23202

tertiary 13301

primary 6851

unknown 1857

Name: education, dtype: int64

default:

no 44396

yes 815

Name: default, dtype: int64

housing:

yes 25130

no 20081

Name: housing, dtype: int64

loan:

no 37967

yes 7244

Name: loan, dtype: int64

contact:

cellular 29285

unknown 13020

telephone 2906

Name: contact, dtype: int64

day:

middle 26186

start 9798

end 9227

Name: day, dtype: int64

month:

may 13766

jul 6895

aug 6247

jun 5341

nov 3970

apr 2932

feb 2649

jan 1403

oct 738

sep 579

mar 477

dec 214

Name: month, dtype: int64

poutcome:

unknown 36959

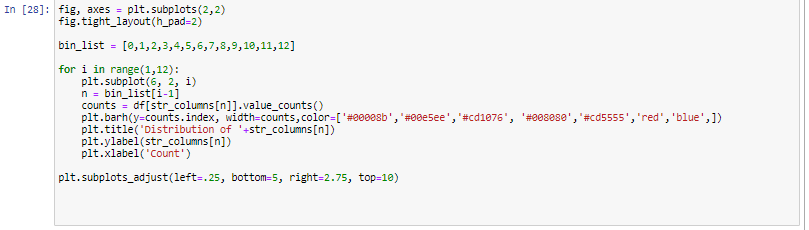
failure 4901

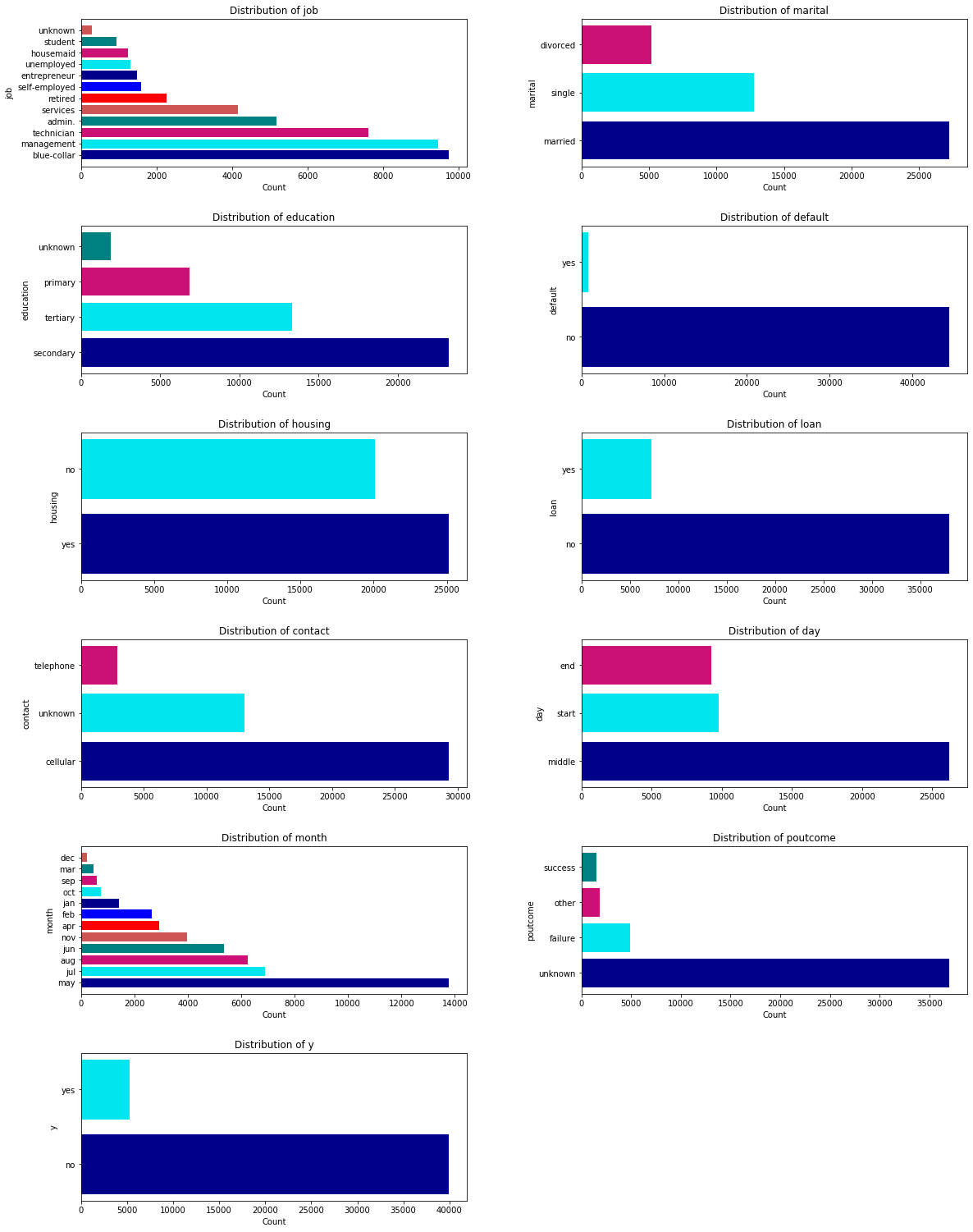
other 1840

success 1511

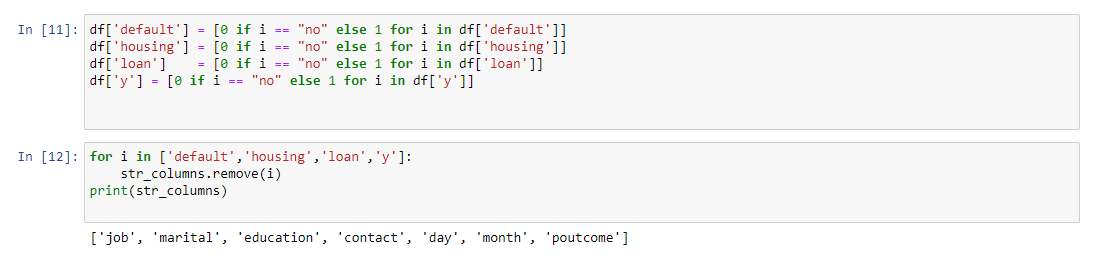
Name: poutcome, dtype: int64

Đồ thị phân tán của từng biến



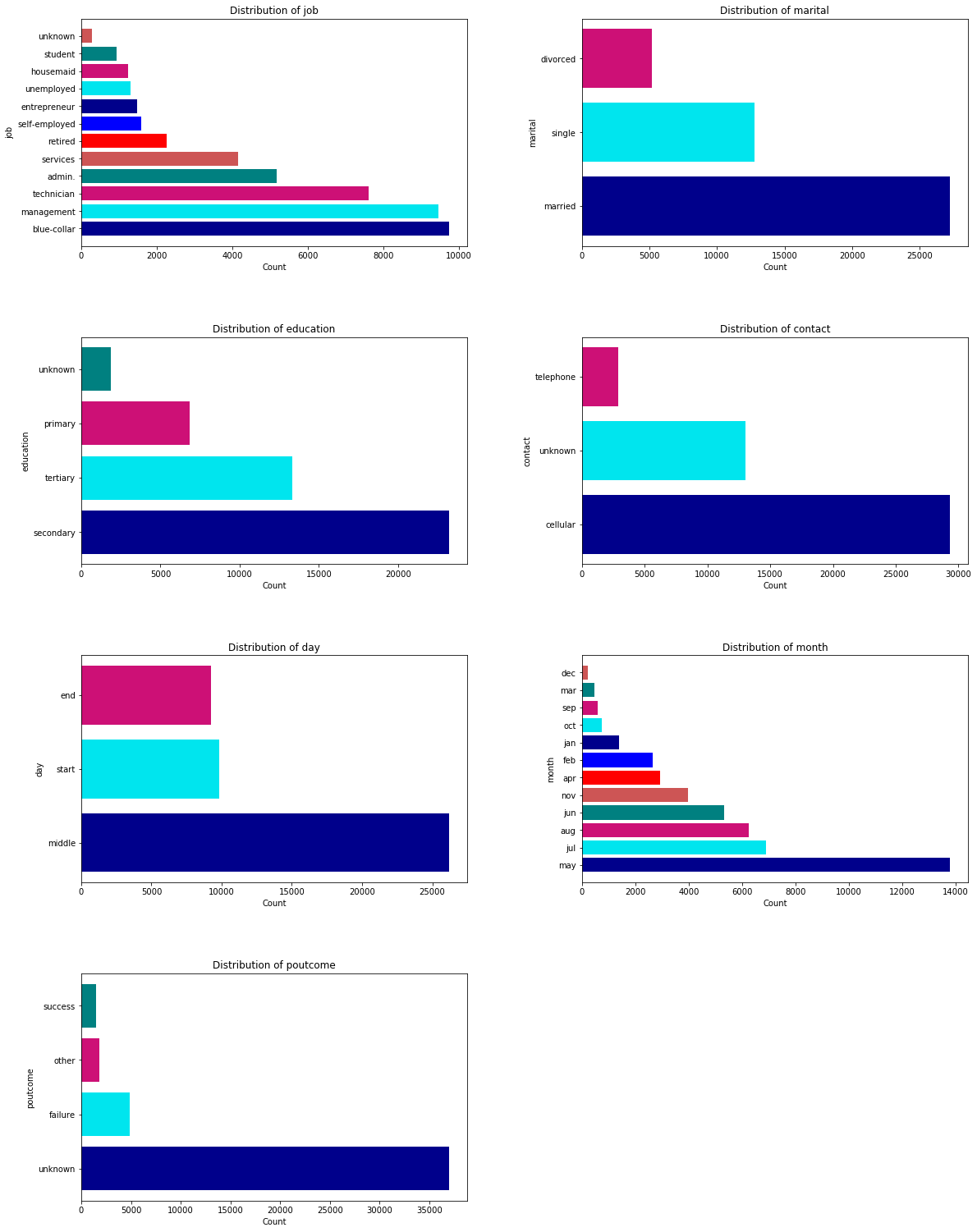


Thay đổi giá trị ký tự trong các cột trên thành 0/1 giá trị nhị phân 0 = "không" và 1 = "có"



Ở trên, chúng tôi đã điều chỉnh các biến phân loại nhị phân và biến chúng thành các biến số mà chúng tôi có thể làm việc với. Bây giờ chúng ta có thể biến đổi các biến phân loại có chứa nhiều hơn một giá trị duy nhất bằng cách sử dụng mã hóa. Trước khi chúng tôi chuyển đổi các biến, chúng tôi phải xem phân phối để có được ý tưởng về hành vi cho mỗi biến





Một số lưu ý từ các bản phân phối trên:

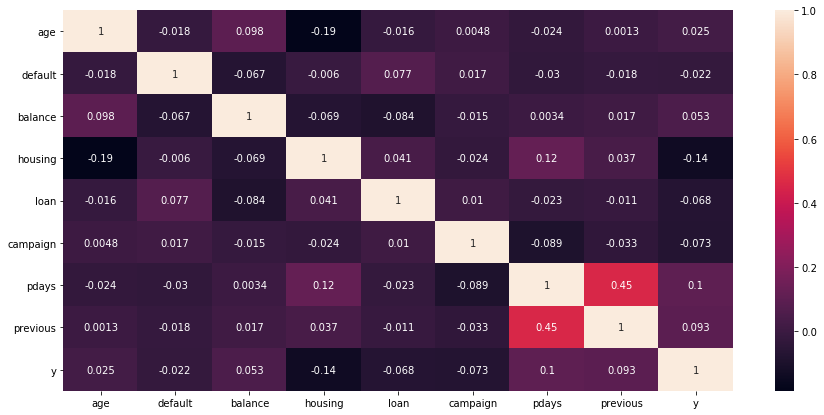
Thật thú vị khi lượng cuộc gọi lớn nhất xảy ra vào những tháng mùa hè và yếu nhất trong những tháng lạnh hơn.

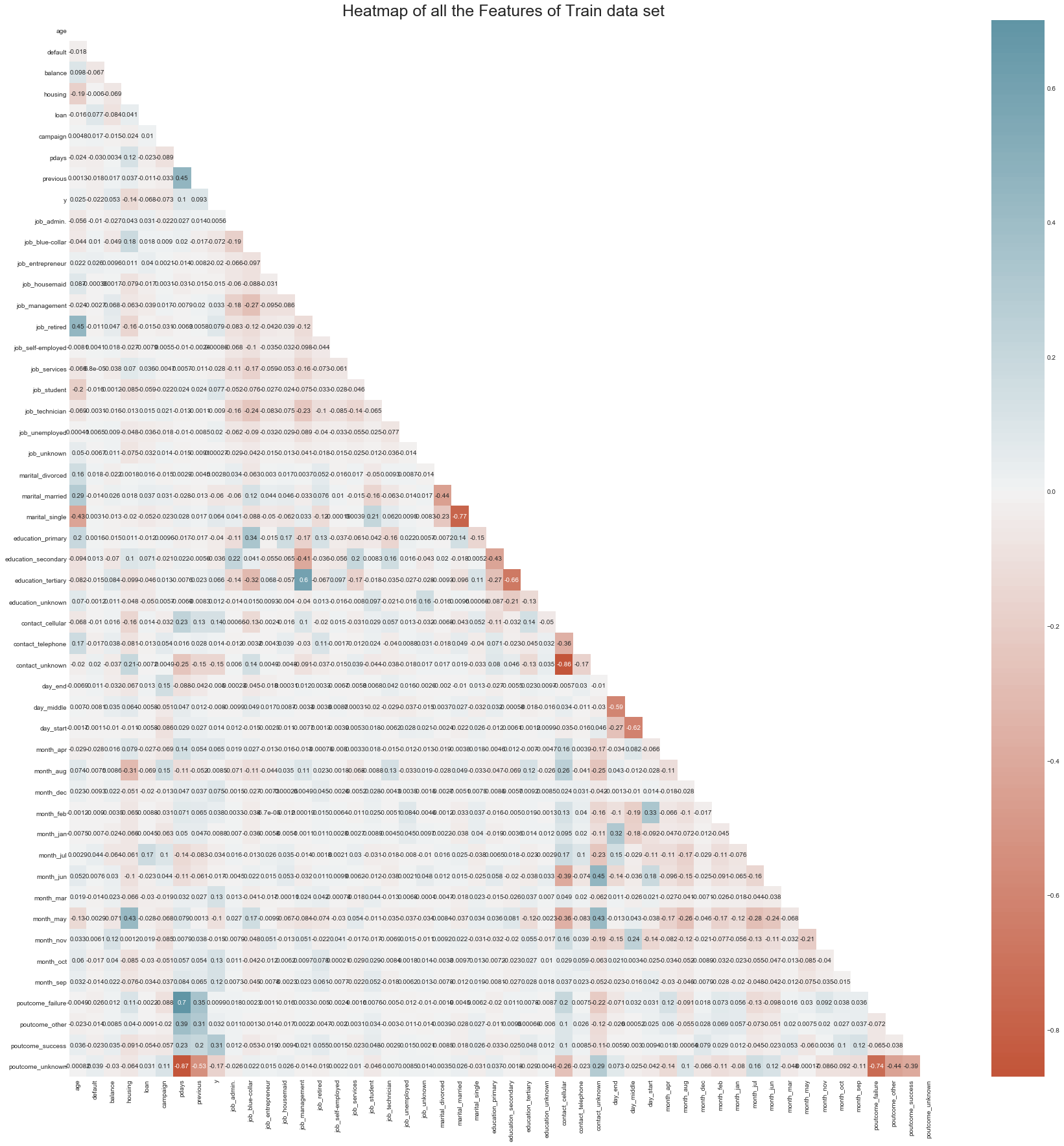
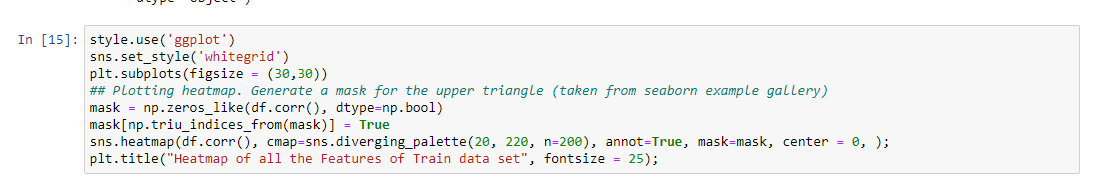
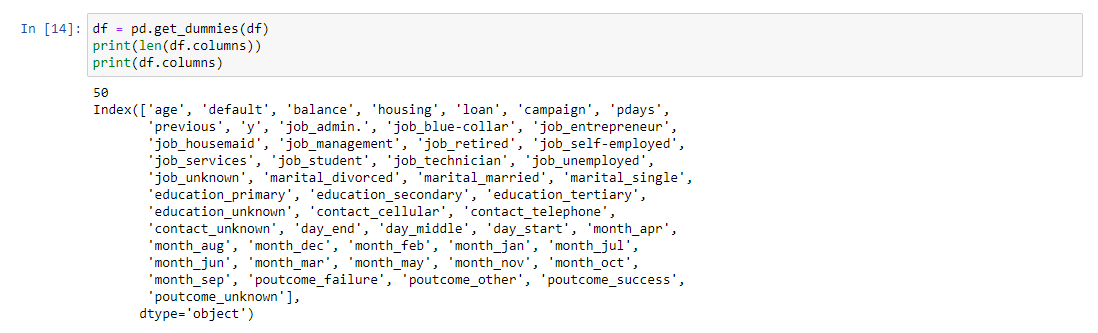
Hãy nhớ rằng "poutcome" phản ánh cách họ phản ứng với chiến dịch trước đó. Chúng tôi cần lưu ý rằng khoảng 75% trong danh mục này được gắn nhãn là "không xác định", phù hợp với những gì chúng tôi đã tìm thấy trước đó.

dựa trên các biến nhân khẩu học, có vẻ như chiến dịch này nhắm mục tiêu đến các cá nhân thuộc tầng lớp trung lưu hỗ trợ thống kê từ biến "số dư" của chúng tôi.

Thông tin này sẽ có giá trị khi chúng tôi xem xét kết quả của mô hình cuối cùng của chúng tôi

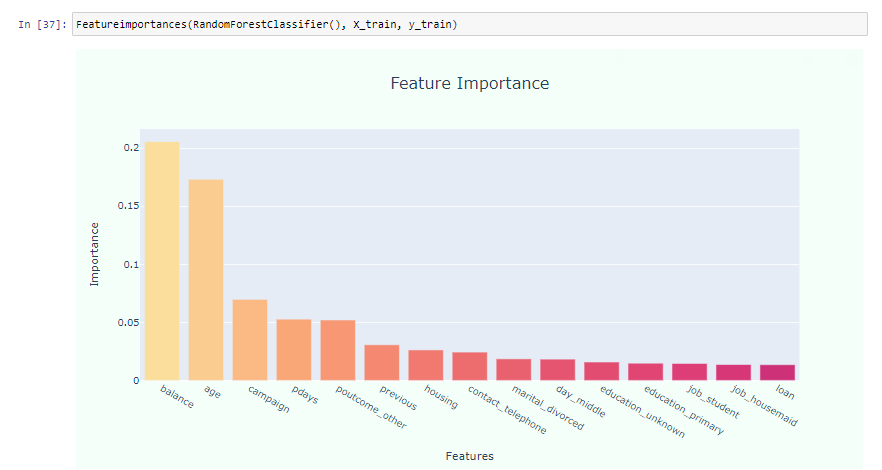
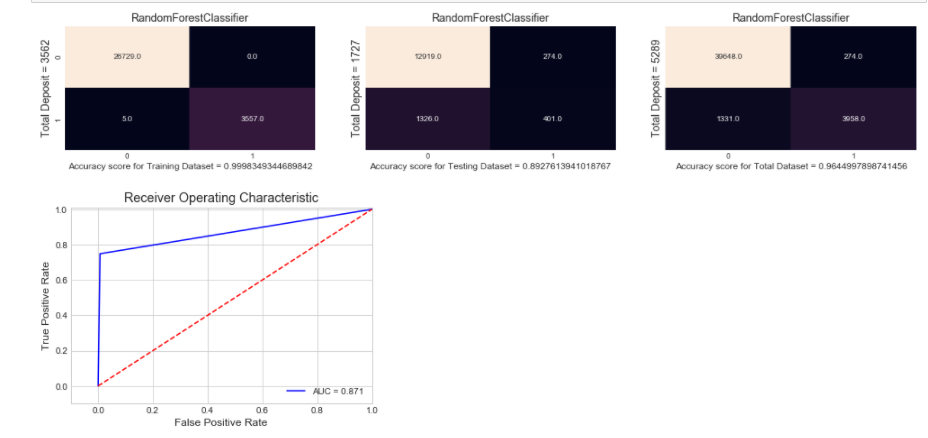
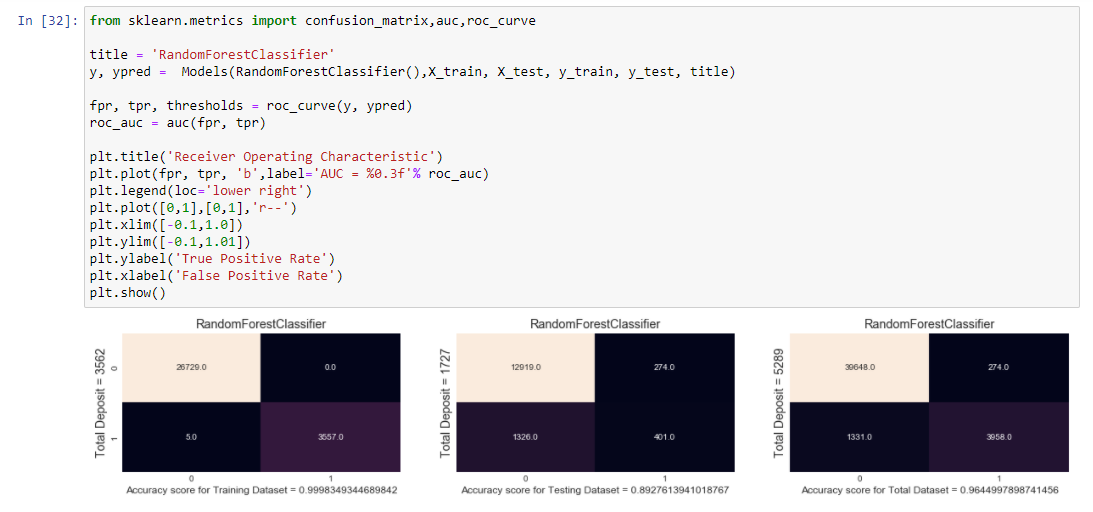
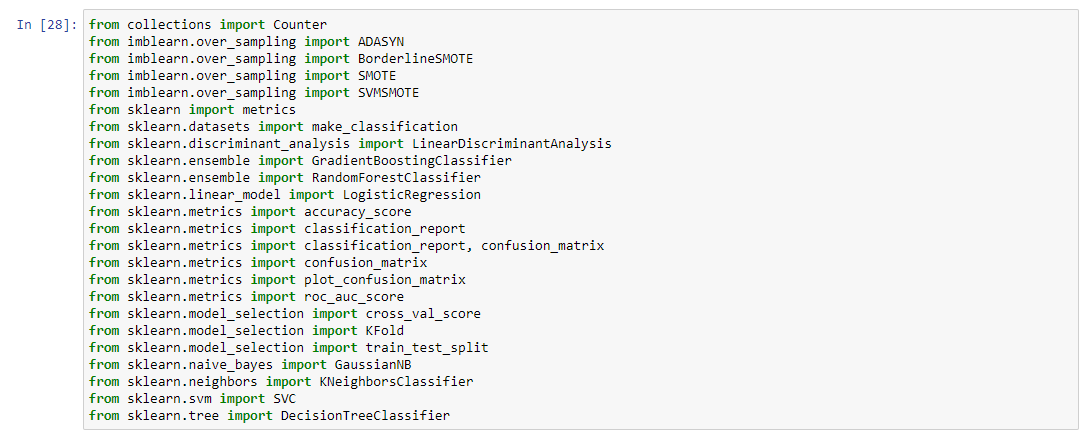
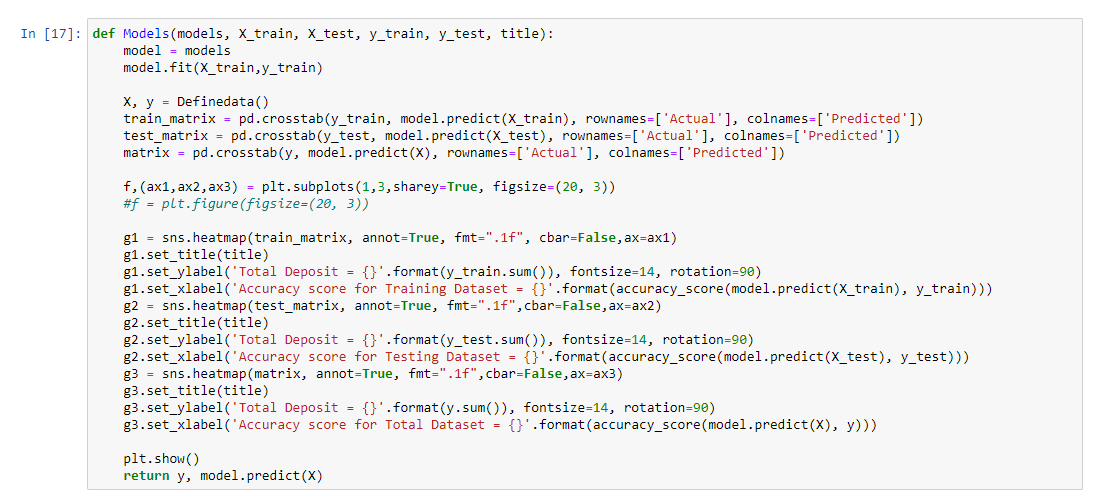
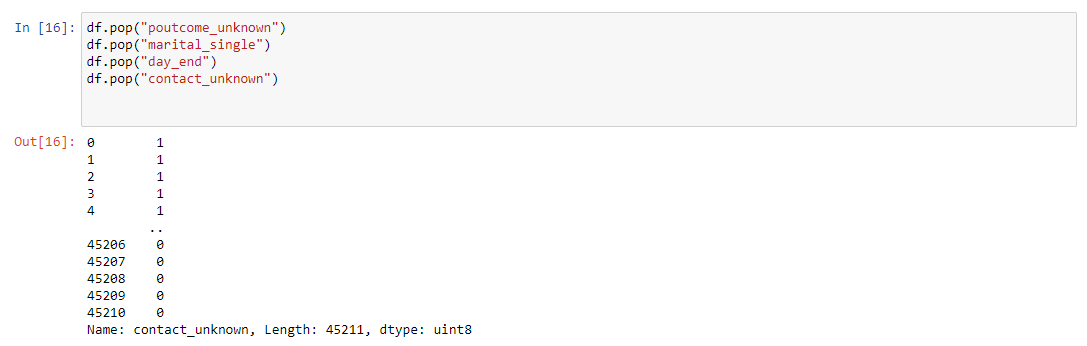
Bảng tương quan của các biến số





Dựa trên bảng tương quan ở trên, các biến phân loại được mã hóa với chính xác 3 giá trị duy nhất dường như có tính đa kim. Điều này có thể là do thực tế là 2 trong số các giá trị được mã hóa sẽ cùng có thông tin về giá trị thứ ba. Ví dụ: nếu có ba lựa chọn giữa pcoucome (thành công, thất bại, không xác định) và nó không phải là thành công hay thất bại, thì dữ liệu sẽ cho rằng nó là không xác định. Do đó, chúng tôi có thể xóa một trong các biến được mã hóa này khỏi mỗi biến thể phân loại với ba giá trị duy nhất.

Xóa: poutcome\_unknown, marital\_single, day\_end, contact\_unknown



Ngay lập tức, chúng tôi có thể chỉ ra rằng hai đặc điểm quan trọng nhất trong mô hình của chúng tôi là sự cân bằng và tuổi tác. Mặc dù biểu đồ này không cho chúng ta biết hướng của đối tượng địa lý (nếu nó ảnh hưởng tích cực hoặc tiêu cực đến kết quả khi mọi thứ khác không đổi), chúng ta có thể ngụ ý một số điều. Thứ nhất, số dư càng lớn thì càng tiết kiệm được nhiều tiền trong ngân hàng. Càng tiết kiệm nhiều tiền trong ngân hàng, bạn càng có nhiều khả năng có thêm tiền để gửi vào một khoản tiền gửi có kỳ hạn. Thứ hai, một người càng lớn tuổi thì thời gian tích lũy của cải càng nhiều. Do đó, họ có nhiều khả năng có thêm tiền để gửi vào một khoản tiền gửi có kỳ hạn. Điều quan trọng có thể là chỉ ra rằng ba tính năng tiếp theo được quy cho chiến dịch. Điều quan trọng cần lưu ý là mặc dù có hai tính năng mạnh mẽ, nhưng có một lượng quan trọng lành mạnh được phân bổ trong số 10 tính năng hàng đầu trong mô hình của chúng tôi.