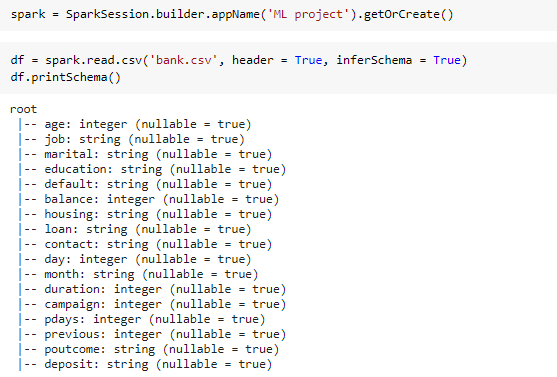
PySpark Kết Hợp Machine Learning

Khi làm việc với Machine Learning trên môi trường PySpark thì ta trải qua các bước như sau:

1. Xử lý tiền dữ liệu
2. Xây dựng mô hình Machine Learning
3. Huấn luyện dữ liệu và đánh giá mô hình

Sau đây là phần code:



Đây là phần code để chúng ta tạo một chương trình PySpark và đọc dữ liệu. Câu lệnh printSchema() giúp ta in ra các header với kiểu dữ liệu nào.

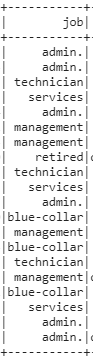
Qua đó ta cũng có thể thấy là có rất nhiều cột có giá trị là String và điều đó khiến ta không thể tính toán nên ta sẽ sử dụng thư viện StringIndexer của PySpark để có thể biến đổi kiểu dữ liệu String sang dạng số.



Cách import thư viện và khai báo thư viện StringIndexer.



Đây là cách sử dụng thì ban đầu ta sẽ cài đặt cho thư viện StringIndexer sẽ nhận vô cột nào và sau khi thực thi việc chuyển đổi thì nó sẽ trả ra cột nào. Sau đó ta chỉ việc đẩy dataset có tên cột mà ta muốn chuyển đổi vô hàm fit() và thư viện sẽ bắt đầu chuyển đổi String qua dạng số qua hàm transform(). Và đây là kết quả trước và sau khi chuyển đổi:

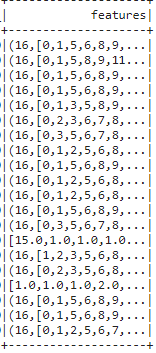
Sau đó ta cần lưu tên các cột thuộc tính và cột kết quả để sau này ta có thể dùng



Sau đó ta cần gôm các cột thuộc tính thành một Vector với dạng mảng 1 chiều để sau này ta có thể dùng một cách dễ dàng hơn. Và thư viện VectorAssembler của pyspark sẽ giúp ta làm việc đó.



Như đã nói ở trên thì ta đã có danh sách tên cột thuộc tính, cho nên ta chỉ cần truyền danh sách đó vô hàm setInputCols() và tên cột mà ta muốn xuất ra sau khi gôm các thuộc tính đó lại là gì qua hàm setOutputCol(). Cuối cùng ta gọi hàm chuyển đổi qua lớp transform và coi kết quả. Và đây là kết quả

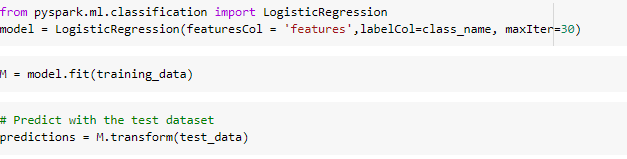


Bước tiếp theo ta sẽ cần tách data ra gồm 2 loại là 1 loại data dùng để test còn loại kia dùng để train.



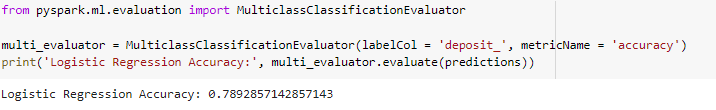
Như đoạn code trên thì dữ liệu train và dữ liệu test được chia với tỷ lệ 8 và 2.

Bước cuối cùng, ta chỉ cần xây dựng Model Machine Learning và đưa các dữ liệu cần thiết vào.



Ở đây ta sẽ dùng thư viện Model Logistic để train dữ liệu và thử kết quả test.

Sau khi ta thực hiện xong thì ta sẽ kiểm tra độ chính xác của thuật toán, thì ta có thể thấy kết quả cho ra độ chính xác lên tới 78,9%.



**Link Tham Khảo:**

<https://www.guru99.com/pyspark-tutorial.html>

https://www.datacamp.com/community/tutorials/apache-spark-tutorial-machine-learning