TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THÀNH PHỐ HỒ CHÍ MINH

KHOA CNTT

ĐỀ THI CUỐI KÌ

Môn thi : Lập Trình Phân Tích Dữ Liệu 1

Lớp/Lớp học phần: _____

Ngày thi: ___/___/_____

Họ và tên sinh viên ……; MSSV:……;

Lưu ý:

- *Không sử dụng USB, chỉ sử dụng tài liệu giấy của cá nhân*
- *Không sử dụng internet, điện thoại, các thiết bị liên lạc*
- *Trình bày nội dung vào giấy thi theo mẫu trên lớp*
- *Sinh viên nộp lại đề thi*
- *Cán bộ coi thi không giải thích gì thêm.*

Đề bài: Công ty phân tích dữ liệu MR. NAM cần phân tích dữ liệu về thị trường ngân hàng. Dữ liệu được thu thập dựa trên các cuộc khảo sát bằng điện thoại đến khách hàng thông qua chiến dịch tiếp thị của viện nghiên cứu tài chính ngân hàng, dữ liệu được lưu trữ trong file **bank-additional-full.csv**. Để tiến hành phân tích dữ liệu, trước tiên công ty cần xác định các đặc tính biến số và thang đo phù hợp. *Dữ liệu được mô tả trong phần phụ lục.*

(**Level 0: 0.5 đ**) Hãy xử lý và tải dữ liệu lên ứng dụng

(**Level 1: 1.5 đ**) Phân tích cần có thống kê số liệu về tình trạng hôn nhân và khoản vay của các khách hàng tham gia khảo sát. Tiến hành trình bày dữ liệu "Martial-Loan".

(**Level 2: 2.0 đ**) Để đạt hiệu quả cao trong việc phân loại xác định tiềm năng của các nhóm cho vay, hãy cung cấp số liệu thống kê và biểu đồ cho phân loại nhóm tuổi của các khách hàng. Ngoài ra người ta cũng cần biểu diễn thông tin các khoản vay ("**loan**") trên từng nhóm tuổi này. Biết rằng người ta phân loại các nhóm tuổi (**plnt**) như sau:

- Từ (-∞, 18): Nhóm tiêu dùng không cho phép vay (viết tắt là: **tdkcp**)
- Từ [18,35): Nhóm tiêu dùng trẻ (viết tắt là: **tdt**)
- Từ [35,50): Nhóm tiêu dùng đẳng cấp (viết tắt là: **tddc**)
- Từ [50,65): Nhóm tiêu dùng an sinh (viết tắt là: **tdas**)
- Từ [65,+∞): Nhóm rủi ro (viết tắt là: **tdrr**)

Hãy trình bày dữ liệu "Age-Loan"

(**Level 3: 3.0 đ**)

Người ta thấy rằng vấn đề trong chiến dịch tiếp thị qua điện thoại thì thời gian trao đổi và tư vấn với khách hàng là một điều quan trọng ảnh hưởng đến chất lượng dữ liệu đầu vào, nhằm quyết định đến chiến lược tung ra các gói sản phẩm nhằm huy động vốn trên các đối tượng phù hợp. Hãy mô tả và khảo sát dạng phân phối của dữ liệu thời gian cuộc gọi (duration) "Duration-Distribution". Bạn là chuyên gia phân tích dữ liệu của công ty MR.NAM, hãy tiến hành xây dựng các kết quả trực quan hóa dữ liệu cho các yêu cầu trên. Điều này hỗ trợ công ty trong quá trình thuyết minh và đàm phán với phía ngân hàng đối tác nhằm đưa ra chiến lược kinh doanh hiệu quả.

---------- **HẾT** ----------

Data Set Information

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Data file "bank-additional-full.csv" with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010)

Attribute Information

Input variables:

 # Bank client data:

1 - age (numeric)

2 - job: type of job (categorical: 'admin', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'selfemployed', 'services', 'student', 'technician', 'unemployed', 'unknown')

3 - marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)

4 - education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')

5 - default: has credit in default? (categorical: 'no', 'yes', 'unknown')

6 - housing: has housing loan? (categorical: 'no', 'yes', 'unknown')

7 - loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

 # Related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular', 'telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day_of_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

 # other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

 # Social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21 - y - has the client subscribed a term deposit? (binary: 'yes', 'no')