ĐỀ THI CUỐI KỲ PHÂN TÍCH DỮ LIỆU 1:

Sử dụng tệp dữ liệu Dethi\_data.xlsx được cung cấp. Sinh viên thực hiện trả lời các yêu cầu bênh dưới trong hệ thống Learning Management System(LMS) của đại học Công Nghiệp

TP.HCM(IUH).

Với câu trả lời điền đáp án thì chỉ điền số cho kết quả và làm tròn 2 chữ số. Ví dụ: 120.00 hoặc 57.89

Với câu chọn đáp án thì chọn câu trả lời đúng nhất, chọn câu trả lời sai sẽ bị trừ điểm.

Giai đoạn 1: tại công ty X, nhóm A tiến hành xây dựng mô hình dự báo giá trị limit thông qua các đặc trunge <c1, d1, f1, c2, d2, f2, f3, f4, f5, f6> bằng mô hình Multip Linear Regression (MLR). Tập dữ liệu ban đầu chia theo hệ số ngẫu nhiên là 10, thành 2 phần tương ứng: tập dữ liệu huyến luyện (df\_train) chiếm 90% và tập dữ liệu kiểm thử (df\_test) chiếm 10%. Nhóm A sử dụng các đại lượng R-Squared và Mean Absolute Error (MAE) để đánh giá mô hình MLR.

Sử dụng mô tả ‘giai đoạn 1’ để giải quyết câu hỏi từ 1,2,3,4,5.

Giai đoạn 2: Sau một thời gian, để cải tiến mô hình MLR, nhóm A đề xuất phát triển mô hình Hybrid Model (HM) để dự báo giá trị limit thông qua các đặc trưng biến đổi Polynomial. Biết rằng việc chia tập dữ liệu cũng giống như ‘giai đoạn 1’ để đánh giá mô hình cũng sử dụng các đại lượng R\_Squared, MAE. Thiết kế của Hybrid Model được mô tả như sơ đồ bên dưới.:

knn\_feature : được trích xuất dựa trên giá trị sác xuất lớn nhất trên nhãn phân lớp (maximum predicted probabilitities for the each) the mô hình K\_Nearest Neighbors KNN

gnb\_feature: được trích xuất dựa trên giá trị xác xuất lớn nhất trên nhãn phân lớp (maximum predicted probalities for the each class) theo mô hình Ggaussian Netive Bayes GNB

các đặc trưng <f3, f4 ,f5, f6> vẫn giữ nguyên giá trị (không biến đổi ).

Các đặt trưng bao gồm <knn\_features\, gnb\_feature, f3, f4, f5, f6> được đưa sang biến đổi đặc trưng Polynomial với degree= 2, có bias và interception.

Sử dụng mô tả ‘giai đoạn 2’ để giải quyết các câu hỏi từ 6 7 8 9, 10

Câu 1: Hãy mô tả tổng quan dữ liệu



a.

Các cột ci, di, fi (với i = 1..6) có 0 dữ liệu thiếu



b.

Cột order có 0 dữ liệu thiếu



c.

Cột delete có 0 dữ liệu thiếu



d.

Cột limit có 0 dữ liệu thiếu



e.

Cột clear có 0 dữ liệu thiếu



f.

Cột remove có 0 dữ liệu thiếu



g.

Kích thước dữ liệu là: (30000, 25)



h.

Cột drop có 0 dữ liệu thiếu

Câu 2: Sau khi tách dữ liệu ra làm 2 phần tập huấn luyện df\_train và tập kiểm thử df\_test . hãy cho biết số lượng mẫu df\_train và df\_test .

a.

Tập dữ liệu df\_test có 3000 mẫu



b.

Tập dữ liệu df\_train có 3000 mẫu



c.

df\_train = (27000,25)



d.

Tập dữ liệu df\_test có 27000 mẫu



e.

Tập dữ liệu df\_train có 27000 mẫu

Câu 3. Sau khi tách dữ liệu ra làm 2 phần tập huấn luyện df\_train và tập kiểm thử df\_test . Hãy cho biết danh sách giá trị order nào thuộc về df\_train, df\_test

a.

[28018, 17729, 29200, 7294, 17674] thuộc về df\_test



b.

[2664, 1392, 18817, 485, 26435] thuộc về df\_test



c.

[20413, 1297, 3907, 20455, 5201] thuộc về df\_train



d.

[28018, 17729, 29200, 7294, 17674] thuộc về df\_train



e.

[20413, 1297, 3907, 20455, 5201] thuộc về df\_test



f.

[24449, 4368, 5750, 13544, 5330] thuộc về df\_test



g.

[2664, 1392, 18817, 485, 26435] thuộc về df\_train



h.

[24449, 4368, 5750, 13544, 5330] thuộc về df\_train

Câu 4: Hãy cho biết giá trị R\_squared và MAE đạt được trên tập dữ liệu kiểm thử theo mô hình MLR

a.

82906.8479776907



b.

106063.21290029172



c.

1.10043969e+04



d.

2.87212966e+04



e.

119691.34323809



f.

6.77891551e-01



g.

3.0035000239325105e-01



h.

1.1568437749190894

Câu 5: hãy cho biết các giá trị đầu vào lần lượt như sau: c1= -1.0, d1 = 9640.0 , f1= 15134.0 , c2 = -2.0 , d2 = 7404.0 , f2 = 0, f3 = 7002.0, f4= 8167.0 , f5= 3996.0, f6 = 2000.0 . Thì giá trị dự báo của limit bởi mô hình MLR là bao nhiêu. (Câu này là câu điền đáp án)

Câu 6: hãy cho biết kết quả của đại lượng độ chính xác (accuracy) trong mô hình trích xuất đặc trưng KNN, với k = 20 trên tập dữ liệu huấn luyện

a.

0.5293



b.

0.529297



c.

0.5292965



d.

0.5292962963



e.

0.529296296296297



f.

0.5292969



g.

0.529296298



h.

0.5292962967

Câu 7: hãy cho biết giá trị trung bình đặc trưng của knn\_feature được trích xuất từ từ tập dữ liệu huấn luyện. (Câu này là câu điền đáp án)

Câu 8: hãy cho biết giá trị nhỏ nhất, trung bình , phương sai, lớn nhất trong ma trận biểu diễn các đặc trưng của tập dữ liệu huấn luyện trước khi đi qua biến đổi đặc trưng Polynomial. Lưu ý: không tính các giá trị thiếu

a.

3372.60711



b.

0.615249



c.

194366041.85558



d.

0.0



e.

1159728.3909



f.

3.73147246



g.

896040.0



h.

1.29371794e-003

Câu 9: hãy cho biết dựa vào giá trị R\_Squared và MAE đạt được trên tập dữ liệu kiểm thử thì mô hình Hybrid và MLR cái nào tốt hơn?

a.

Mô hình MLR tốt hơn



b.

Mô hình Hybrid và MLR tương đương nhau



c.

R-Squared của Hybrid Model là 0.1923326662875111



d.

Sai số đánh giá MAE trên mô hình Hybrid nhỏ hơn MLR



e.

MAE của Hybrid Model là 6063.2127933



f.

R-Squared của Hybrid Model là 0.57437749190894



g.

MAE của Hybrid Model là 82906.8629172



h.

MAE của Hybrid Model là 91741.45769423115



i.

Sai số đánh giá MAE trên mô hình Hybrid lớn hơn MLR



j.

R-Squared của Hybrid Model là 0.3604978176



k.

Mô hình Hybrid tốt hơn

Câu 10: Hãy cho biết các giá trị đầu vào lần lượt như sau: c1 = -1.0, d1 = 9640, f1=15134, c2 = -2.0, d2=7404, f2=0, f3=7002, f4=8167, f5=3996, f6=2000. Thì giá trị dự báo của limit bởi mô hình HM sẽ là bao nhiêu? (Câu này là câu điền đáp án)