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 phatys triển mô hình Hybrid Model (HM) để dự báo giá trị limit thông qua các đặc trưng <knn_feature, gnb_feature, f3, f4, f5 ,f6 >. Biết rằng việc chia tập dữ liệu cũng giống như giai đoanh 1 để đánh giá mô hình cũng sử dụng các đại lượng R_Squared, MED. 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 (mãimum 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). 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

CÂU 1

a.

Cột drop có 0 dữ liệu thiếu

b.

Kích thước dữ liệu là: (30000, 25)

c.

Cột limit có 0 dữ liệu thiếu

d.

Cột clear có 0 dữ liệu thiếu

e.

Cột delete có 0 dữ liệu thiếu

f.

Cột remove có 0 dữ liệu thiếu

g.

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

h.

Cột order có 0 dữ liệu thiếu

câu 2: sau khi tách dữ liệu ra làm 2 phần tập huấn luyện df_train và df_test . hãy cho biết số lượng mẫu df_train và df_test .

Tập dữ liệu df_train có 27000 mẫu

b.

Tập dữ liệu df_test có 3000 mẫu

c.

Tập dữ liệu df_test có 27000 mẫu

d.

Tập dữ liệu df_train có 3000 mẫu

e.

df_train = (27000,25)

3. sau khi chia tách dữ liệu thành 2 phần tập dữ liệu tập huấn là train, tập kiểm thử là df_test. Hãy cho biết danh sách giá trị order nào thuộc về df_train, df_test

a.

[2664, 1392, 18817, 485, 26435] thuộc về df_train

b.

[28018, 17729, 29200, 7294, 17674] thuộc về df_test

c.

[2664, 1392, 18817, 485, 26435] thuộc về df_test

d.

[24449, 4368, 5750, 13544, 5330] thuộc về df_test

e.

[20413, 1297, 3907, 20455, 5201] thuộc về df_test

f.

[28018, 17729, 29200, 7294, 17674] thuộc về df_train

g.

[20413, 1297, 3907, 20455, 5201] thuộc về df_train

h.

[24449, 4368, 5750, 13544, 5330] thuộc về df_train

câu 4: dùng giai đoạn 1 hãy cho biết giá trị R_squared và MED đạt được trên tập dữ liệu kiểm thử theo mô hình MLR CÂU 4

a.

2.87212966e+04

b.

106063.21290029172

c.

119691.34323809

d.

82906.8479776907

e.

3.0035000239325105e-01

f.

1.1568437749190894

g.

6.77891551e-01

h.

1.10043969e+04

Câu 5: hãy cho biết các gái trị dầu 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.

Hãy làm tròn đáp án với 5 chữ số thập phân

Không được để khoảng cách thừa

Ví dụ: **12389456.09087**

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.

0.5292965

b.

0.5292969

c.

0.529297

d.

0.5292962967

e.

0.5292962963

f.

0.529296296296297

g.

0.529296298

h.

0.5293

Câu 7: hãy cho biết giá trị trung bình đặc trưng của knn_feature được trích xuất từ từ tập dữ liệu huấn luyên

Hãy làm tròn đáp án với 10 chữ số thập phân

Không được để khoảng cách thừa

Ví dụ: **12389456.0908700001**

Câu 8: hayc cho biết giá tị nhior nhất, trung bình , phương sai, lớn nhất trong ma trận biểu diễn các đặc trưng của taapj dữ liệu guấn luyện, lưu ý không tính các giá trị thiếu

a.

1.29371794e-003

b.

0.615249

c.

1159728.3909

d.

896040.0

e.

194366041.85558

f.

3372.60711

g.

0.0

h.

3.73147246

Câu 9: hãy cho biết giá trị R_Squared và MED đạt được trên tập dữ liệu kiểm thử thì mô hình Hybrid và MLR cái nào tốt hơn?

a.

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

b.

Mô hình Hybrid tốt hơn

c.

R-Squared của Hybrid Model là 0.3604978176

d.

MAE của Hybrid Model là 96495.0545868644

e.

Mô hình Hybrid và MLR tương đương nhau

f.

R-Squared của Hybrid Model là 0.57437749190894

g.

MAE của Hybrid Model là 82906.8629172

h.

MAE của Hybrid Model là 6063.2127933

i.

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

j.

R-Squared của Hybrid Model là 0.1274168005321744

k.

Mô hình MLR tốt hơn