\* Phân tích các yếu tố ảnh hưởng đến việc hài lòng của khách hàng

\*\* Đổi kiểu dữ liệu của các biến phân loại thành integer trước khi đưa vào mô hình hồi quy

> new\_data\_2$nearest\_warehouse = factor((new\_data\_2$nearest\_warehouse), levels = c("Bakers","Nickolson","Thompson"))

> new\_data\_2$season = factor((new\_data\_2$season), levels = c("Spring", "Summer", "Autumn","Winter"))

> new\_data\_3 = new\_data\_2[,c("nearest\_warehouse","delivery\_charges","customer\_lat","customer\_long","coupon\_discount", "season", "order\_total", "distance\_to\_nearest\_warehouse", "is\_expedited\_delivery", "is\_happy\_customer")]

> new\_data\_3$season = as.integer(new\_data\_3$season)

> new\_data\_3$nearest\_warehouse = as.integer(new\_data\_3$nearest\_warehouse)

\*\* Xét mô hình hồi quy logistic gồm biến is\_happy\_customer là biến phụ thuộc và các biến còn lại là biến độc lập

> mohinh1 = glm(is\_happy\_customer~., family = "binomial", data = new\_data\_3)

> summary(mohinh1)

A screenshot of a computer

Description automatically generated

-> Các biến như nearest\_warehouse, customer\_long, coupon\_discount, order\_total có p-value > mức ý nghĩa 5% nên không có giá trị thống kê, còn biến customer\_lat cũng không có giá trị thống kê vì chỉ mỗi 1 trị số trong tọa độ vị trí của khách hàng.

Xét mô hình hồi quy mới sau khi loại bỏ các biến không có giá trị thống kê trên

> new\_data\_4 = new\_data\_3[,c("delivery\_charges", "season", "distance\_to\_nearest\_warehouse", "is\_expedited\_delivery", "is\_happy\_customer")]

> mohinh2 = glm(is\_happy\_customer~., family = "binomial", data = new\_data\_4)

> summary(mohinh2)

A screenshot of a computer

Description automatically generated

-> loại bỏ những biến không có giá trị thống kê trong mô hình 1 cho mô hình 2 chỉ gồm các biến có giá trị thống kê. Giá trị AIC của 2 mô hình gần như tương đương nhau

\* Kiểm định mô hình bằng ANOVA

Giả thuyết H0: 2 mô hình có hiệu quả như nhau.

Giả thuyết H1: 2 mô hình có hiệu quả khác nhau.

> anova(mohinh1, mohinh2, test = "LRT") (LRT = Likelihood Ratio Test)

A computer screen shot of a program

Description automatically generated

-> Từ kết quả phân tích trên ta thấy p-value = 0.06775 > 0.05 nên ta không có cơ sở để bác bỏ H0. Ở mức ý nghĩa 5%, chúng ta có thể kết luận rằng cả 2 mô hình là tương tự nhau nhưng ta chọn mô hình 2 vì mô hình này có chứa các biến có ý nghĩa thống kê.

Ta tính được logit(P) (với P là xác suất khách hàng hài lòng):

Logit(P) = -12.36417 + 0.25483.delivery\_charges – 0.54349.season – 1.48004.distance\_to\_nearest\_warehouse -4.16165.is\_expedited\_delivery.

Ý nghĩa các tham số trong mô hình hồi quy:

Hệ số chặn (intercept)= -12.36417 là giá trị logit(P) khi các biến độc lập còn lại bằng 0. Trong trường hợp này logit(P)= -12.36417.

Null deviance: Độ lệch khi không có biến độc lập trong mô hình.

Residual deviance: Độ lệch khi có biến độc lập trong mô hình

Hệ số hồi quy của biến độc lập vừa phản ánh mức độ tác động đồng thời cũng thể hiện chiều tác động của biến độc lập lên biến phụ thuộc. Ví dụ: đối với chi phí giao hàng (delivery\_charges), nếu tăng chi phí lên 1 đơn vị thì giá trị của logit(P) tăng lên 0.25483, điều này ảnh hưởng đến mức độ hài lòng của khách hàng.