multicollinearity : avg\_ratings vs rate\_5 4 3 2 1

1. Random Forest Regression

Metrics sử dụng: Mean Absolute Error (MAE) hoặc Soft Interval Accuracy (SIA)

#category: categorical

#price: numerical

#shop\_address: categorical

#avg\_rating: numerical

#n\_reviews: numerical

#n\_loved: numerical

#shop\_n\_review: numerical

#shop\_n\_product: numerical

#shop\_rate\_feedback: numerical

#shop\_time\_feedback: numerical

#shop\_age: numerical

#shop\_follower: numerical

Train - Val - Test: 60 - 20- 20

1. **Dự đoán sử dụng toàn bộ features (one hot + numeric): 82 chiều**

Kết quả:

MAE train: 159.64

MAE val: 413.99

MAE test: 472.69

Ghi chú: do kết quả dự đoán có thể lên đến đơn vị nghìn (lượt bán), để dễ giải thích, ta sẽ sử dụng độ đo MAE (thay vì MSE như thông thường). Do đó ta có thể đánh giá MAE trên tập test ~473 có nghĩa là mô hình dự đoán trung bình sai số khoảng 473 cho một sản phẩm trong tập test.

Bây giờ ta sẽ xét SIA metric (metric này có nghĩa là cho trước 1 giá trị epsilon, một dự đoán được xem là đúng (True) khi nó có giá trị nằm trong khoảng [y - epsilon, y + epislon] với y là groundtruth label).

Với epsilon = 1000 (cho phép sai số dự đoán lên đến 1k lượt bán):

SIA train: 0.975

SIA val: 0.919

SIA test: 0.921

epsilon = 500 (cho phép sai số dự đoán lên đến 500 lượt bán):

SIA train: 0.944

SIA val: 0.843

SIA test: 0.835

epsilon = 200 (cho phép sai số dự đoán lên đến 200 lượt bán):

SIA train: 0.861

SIA val: 0.67

SIA test: 0.67

Nhận xét: ta thấy mô hình đạt được độ chính xác chấp nhận được. Hiệu năng trên tập val và test xấp xỉ nhau, trái ngược với khi sử dụng MSE metric, tập test có MSE lớn hơn rất nhiều so với val. Điều này có thể là do có một số outlier làm lệch kết quả dự đoán.

1. **Chỉ sử dụng các numerical features (10 chiều):**

price, avg\_rating, n\_reviews,n\_loved, shop\_n\_review, shop\_n\_product, shop\_rate\_feedback, shop\_time\_feedback, shop\_age, shop\_follower

(Bỏ đi category và shop\_address)

Kết quả:

MAE train: 169.52

MAE val: 433.01

MAE test: 491.87

epsilon = 1000:

SIA train: 0.975

SIA val: 0.916

SIA test: 0.909

epsilon = 500:

SIA train: 0.942

SIA val: 0.835

SIA test: 0.825

epsilon = 200:

SIA train: 0.853

SIA val: 0.641

SIA test: 0.655

Nhận xét: ta thấy mô hình đạt được kết quả thấp hơn một ít so với khi sử dụng thêm feature category và shop\_address (Điều này dễ hiểu vì loại sản phẩm và nơi bán cũng ảnh hưởng đến số lượng bán ra).

1. **Chỉ sử dụng các features liên quan đến shop và sản phẩm (loại bỏ các feature về rating, số lượt yêu thích,...)**

category, price, shop\_address, shop\_n\_review, shop\_n\_product, shop\_rate\_feedback, shop\_time\_feedback, shop\_age, shop\_follower

(Bỏ avg\_rating, n\_reviews, n\_loved)

Kết quả:

MAE train: 317.349

MAE val: 805.426

MAE test: 906.455

epsilon = 1000:

SIA train: 0.943

SIA val: 0.824

SIA test: 0.831

epsilon = 500:

SIA train: 0.877

SIA val: 0.700

SIA test: 0.708

epsilon = 200:

SIA train: 0.723

SIA val: 0.492

SIA test: 0.490

Nhận xét: Ta thấy kết quả dự đoán có độ chính xác giảm mạnh so với 2 thử nghiệm a và b (có sử dụng feature liên quan đến rating, yêu thích). Do đó ta có thể kết luận các feature này đóng vai trò rất quan trọng trong việc dự đoán chính xác.

1. Linear Regression
   1. **Dự đoán sử dụng toàn bộ features (one hot + numeric): 82 chiều**

Trước khi huấn luyện mô hình, ta sẽ chuẩn hóa các features theo phương pháp Z-score normalization.

Kết quả đạt được:

MAE train: 615.818

MAE val: 548840935861.9957

MAE test: 433296680921.9505

SIA với epsilon = 500:

SIA train: 0.721

SIA val: 0.702

SIA test: 0.708

Nhận xét: Ta thấy mô hình có SIA thấp hơn so với khi sử dụng Random Forest khá nhiều. Hơn nữa, ta thấy mô hình có MAE trên tập val và tập test cao bất thường. Sau khi tìm hiểu nguyên nhân thì ta thấy có một vài hệ số (coefficient) của một số feature có giá trị rất cao (Ví dụ như 8.94728651e+14). Hơn nữa, ta nhận thấy các hệ số này thuộc về feature shop\_address (onehot). Nguyên nhân có thể là do việc tối ưu (Ordinary Least Square) và nghịch đảo một ma trận thưa (sparse matrix). Để khắc phục, ta sẽ sử dụng L2 Regularization (Ridge) để giới hạn giá trị của các hệ số.

Kết quả khi sử dụng Ridge (tìm hyperparameter alpha tốt nhất sử dụng phương pháp gridsearch trên tập val), ta có:

Best cross-validation MAE: 582.711

with alpha: 10.0

MAE train: 609.247

MAE val: 582.710

MAE test: 615.954

Sau khi kiểm tra các hệ số, ta không thấy các giá trị bất thường như ở trên.

SIA với epsilon = 500:

SIA train: 0.721

SIA val: 0.704

SIA test: 0.707

Ta thấy độ chính xác vẫn thấp hơn phương pháp Random Forest khá nhiều (điều này có thể là do mô hình vẫn chưa đủ phức tạp, đặc biệt là đối với các categorical feature chăng?)

* 1. **Linear Regression chỉ sử dụng các numerical features**

Kết quả:

MAE train: 600.929

MAE val: 567.050

MAE test: 612.985

SIA với epsilon = 500:

SIA train: 0.764

SIA val: 0.769

SIA test: 0.745

Nhận xét: ta thấy kết quả có phần cao hơn so với khi sử dụng các categorical features. Điều này có thể là do mô hình linear regression chưa phù hợp với dữ liệu dạng này (có thể có các giá trị category chỉ xuất hiện trong tập test nhưng không xuất hiện trong tập train). Tiếp theo ta sẽ thử sử dụng Mean Encoding thay thế cho One Hot Encoding (đối với categorical feature)

* 1. **Linear Regression sử dụng Mean Encoding cho các categorical features**

Kết quả:

MAE train: 602.964

MAE val: 565.976

MAE test: 611.147

SIA với epsilon = 500:

SIA train: 0.755

SIA val: 0.765

SIA test: 0.743

Ta thấy kết quả có phần giảm đi đôi chút so với khi chỉ sử dụng các numerical feature có sẵn. Ngoài ra, ta thấy kết quả trên tập train cũng xấp xỉ kết quả trên tập val và test (ngược lại với mô hình Random Forest có độ lỗi trên tập train thấp hơn rất nhiều so với val và test). Có thể là do mô hình Linear Regression chưa phải là phù hợp nhất với tập dữ liệu của ta (mô hình có thể đưa ra kết quả dự đoán mang giá trị âm).

1. **Neural Network Regression**
   1. **Sử dụng toàn bộ features (onehot encoding + numeric)**

Tương tự như linear regression , ta sẽ tiền xử lý input bằng việc chuẩn hóa các features theo phương pháp Z-score normalization.

Ta sẽ thực nghiệm với 1 mạng Neural Network nhỏ như sau:

***Input Vector -> FC(32 units) + ReLU -> FC(64 units) + ReLU -> Output (1 unit)(+ReLU)***

Kết hợp với L2 Regularization (alpha=0.01) và Hàm mất mát MAE.

Kết quả sau khi huấn luyện (ta chọn bộ trọng số có giá trị MAE val thấp nhất để đánh giá):

MAE train: 405.848

MAE val: 391.498

MAE test: 428.935

SIA với epsilon = 500:

SIA train: 0.867

SIA val: 0.852

SIA test: 0.843

* 1. **Chỉ sử dụng các numerical features**

Kết quả:

MAE train: 449.249

MAE val: 414.424

MAE test: 474.167

SIA với epsilon = 500:

SIA train: 0.840

SIA val: 0.836

SIA test: 0.823

Ta thấy kết quả của mô hình cũng có phần giảm đi so với khi sử dụng toàn bộ các features (Tương tự như phần Random Forest)

* 1. **sử dụng Mean Encoding cho các categorical features**

Kết quả:

MAE train: 433.238

MAE val: 401.349

MAE test: 457.555

SIA với epsilon = 500:

SIA train: 0.851

SIA val: 0.848

SIA test: 0.836

**Kết Luận:** thực nghiệm cho thấy 2 mô hình cho kết quả tốt nhất là Random Forest và Neural Network với toàn bộ các feature (onehot encoding + numerical). Tuy nhiên mô hình Neural Network có thời gian huấn luyện khá lâu so với Random Forest. Do đó nhóm sẽ chọn mô hình Random Forest để làm demo ứng dụng.