Trường ĐH Khoa Học Tự Nhiên Tp. Hồ Chí Minh
TRUNG TÂM TIN HỌC

Đồ án tốt nghiệp Data Science

Topic: ***Sentiment Analysis***

GVHD: Cô Khuất Thùy Phương

HVTH: Đỗ Thị Phượng

2022

Nội dung

1. **Project Overview**
2. **EDA**
3. **Sentiment Analysis – Logistic Regression Model**
4. **Prediction**

Project Overview

Business Objective/Problem

- Foody.vn là một kênh phối hợp với các nhà hàng/quán ăn bán thực phẩm online.
- Chúng ta có thể lên đây để xem các đánh giá, nhận xét cũng như đặt mua thực phẩm.
- Từ những đánh giá của khách hàng, vấn đề được đưa ra là làm sao để các nhà hàng/ quán ăn hiểu được khách hàng rõ hơn, biết họ đánh giá về mình như thế nào để cải thiện hơn trong dịch vụ/ sản phẩm.

Nội dung

1. Project Overview
2. EDA
3. Sentiment Analysis – Logistic Regression Model
4. Prediction

EDA

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22734 entries, 0 to 22733
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   restaurant    22734 non-null  object 
 1   review_text   22734 non-null  object 
 2   review_score  22734 non-null  float64
dtypes: float64(1), object(2)
memory usage: 533.0+ KB
```

```
data = data.drop_duplicates()
data.shape
```

```
(22734, 3)
```

```
data[data.isna().T.any()]
```

restaurant	review_text	review_score

```
data[data.isnull().T.any()]
```

restaurant	review_text	review_score

```
len(data.restaurant.unique())
```

```
567
```

- Dữ liệu do học viên tự cào trên foody.vn với các nhà hàng ở Hà Nội, TP.HCM, Thanh Hóa, Cần Thơ, Vũng Tàu… với tất cả các loại món ăn, đồ uống…
- Gồm 22734 records (reviews) cho 567 nhà hàng.
- Không có dữ liệu trùng/NaN/null

EDA

- Hiển thị một vài dòng dữ liệu:

```
data.head()
```

	restaurant	review_text	review_score
0	Chả Cá Hà Nội Xưa	Nhà hàng mới đổi địa chỉ sang 24 Hồng Hà, khá ...	8.8
1	Chả Cá Hà Nội Xưa	Quán đã chuyển về 22 Hồng Hà cới phong cách kh...	3.2
2	Chả Cá Hà Nội Xưa	Giá niêm yết trên foody một kiểu, giá lúc ship...	3.2
3	Chả Cá Hà Nội Xưa	Xem review thấy mọi người khen chả cá lã vọng ...	7.0
4	Chả Cá Hà Nội Xưa	Tối nay mới đi ăn quán này món chả cá lã vọng ...	9.2

```
data.tail()
```

	restaurant	review_text	review_score
22729	Rodstarz FastFood Restaurant - Trương Công Định	Hôm nay trời mưa hội bạn của mình rủ đi ăn món...	7.0
22730	Rodstarz FastFood Restaurant - Trương Công Định	Lâu lâu chạy ngang thấy quán có món ăn mới nên...	7.0
22731	Rodstarz FastFood Restaurant - Trương Công Định	Vị trí thuận lợi dễ tìm, tuy vậy không gian kh...	7.8
22732	Rodstarz FastFood Restaurant - Trương Công Định	Vào "ăn sáng" nhưng được giá cơm trưa. Chị nhâ...	8.4
22733	Rodstarz FastFood Restaurant - Trương Công Định	Địa điểm quen thuộc của nhóm mình, bọn mình ăn...	7.0

EDA

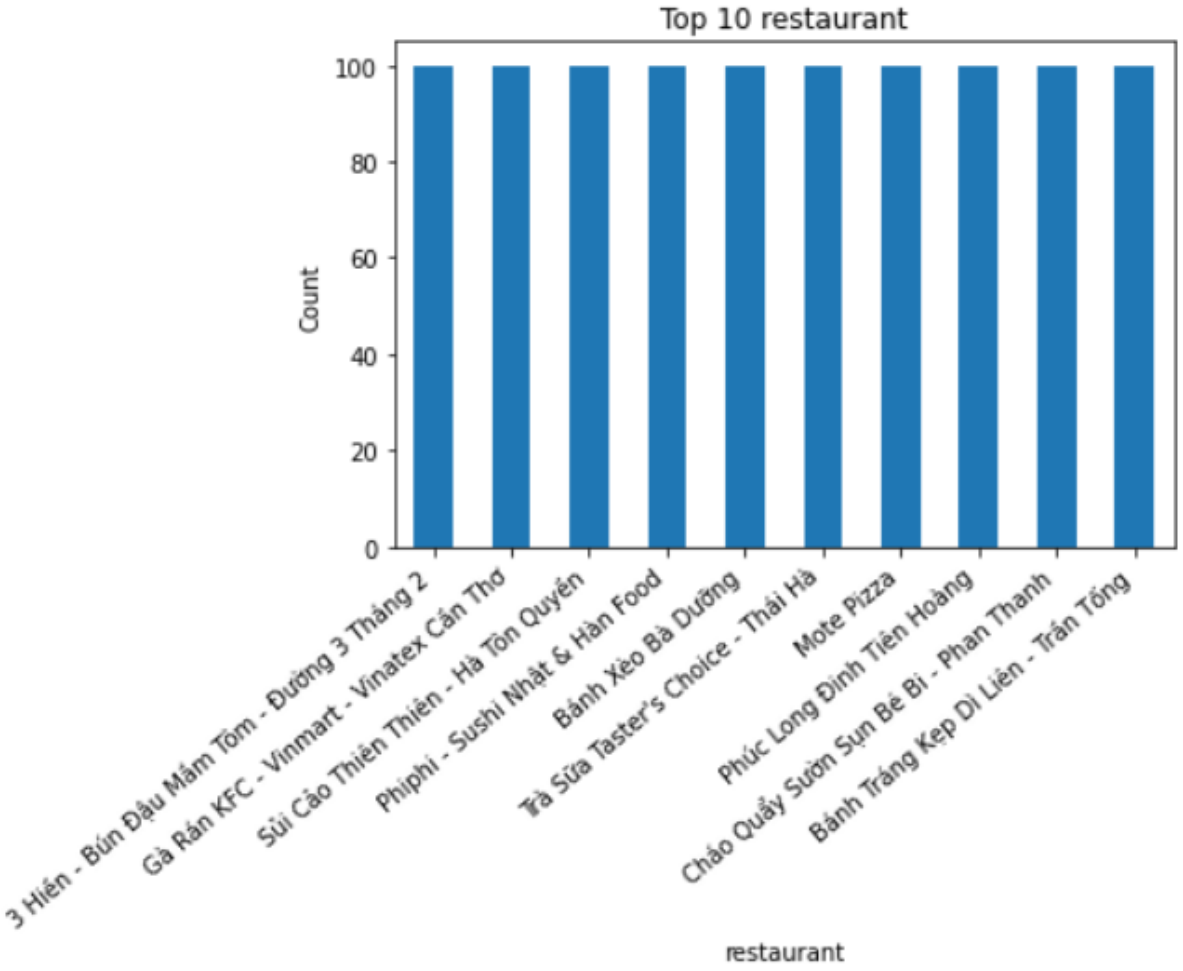
- Một số nhà hàng có nhiều reviews nhất là 100 reviews. Một số nhà hàng chỉ có 1 review.

```
restaurant_ = data.groupby('restaurant')['review_text'].count().sort_values(ascending = False)
restaurant_
```

```
restaurant
3 Hiền - Bún Đậu Mắm Tôm - Đường 3 Tháng 2                  100
Gà Rán KFC - Vinmart - Vinatex Cần Thơ                      100
Sủi Cảo Thiên Thiên - Hà Tôn Quyền                          100
Phiphi - Sushi Nhật & Hàn Food                              100
Bánh Xèo Bà Dưỡng                                           100
                                                           ...
Trà Sữa House Of Cha - Trần Hữu Dực                           1
Family Chicken - Gà Rán, Bánh Gà & Đồ Ăn Vặt - Chùa Quỳnh     1
Trà Sữa Mixue - Xuân Thủy                                     1
GuChi - Tokbokki, KimBap & Cơm Trộn - Phạm Văn Đồng           1
Tocotoco - Lĩnh Nam                                           1
Name: review_text, Length: 567, dtype: int64
```

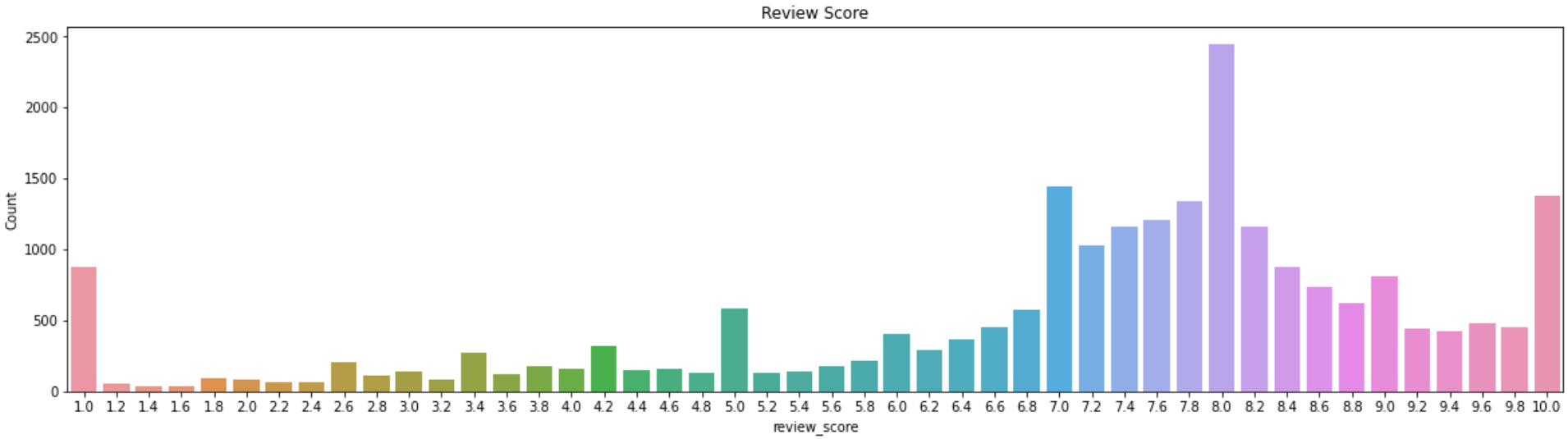
EDA

- Top 10 nhà hàng có nhiều reviews nhất:



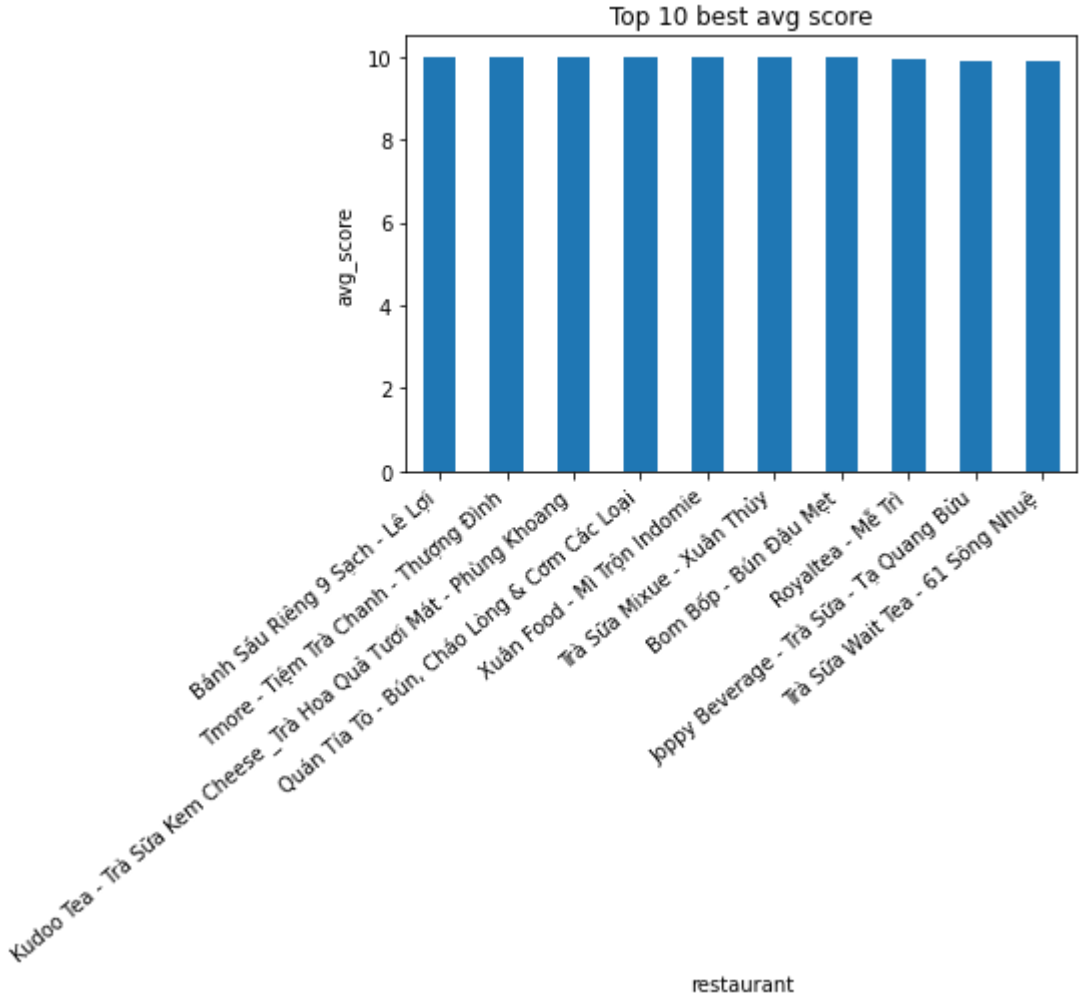
EDA

- Hầu hết người dùng cho điểm khá cao từ 7.0 trở lên. Nhiều nhất ở mức 8.0 điểm. Ngoài ra có một số mốc đáng chú ý như 1.0; 5.0; 7.0, 8.0, 9.0 và 10.0.



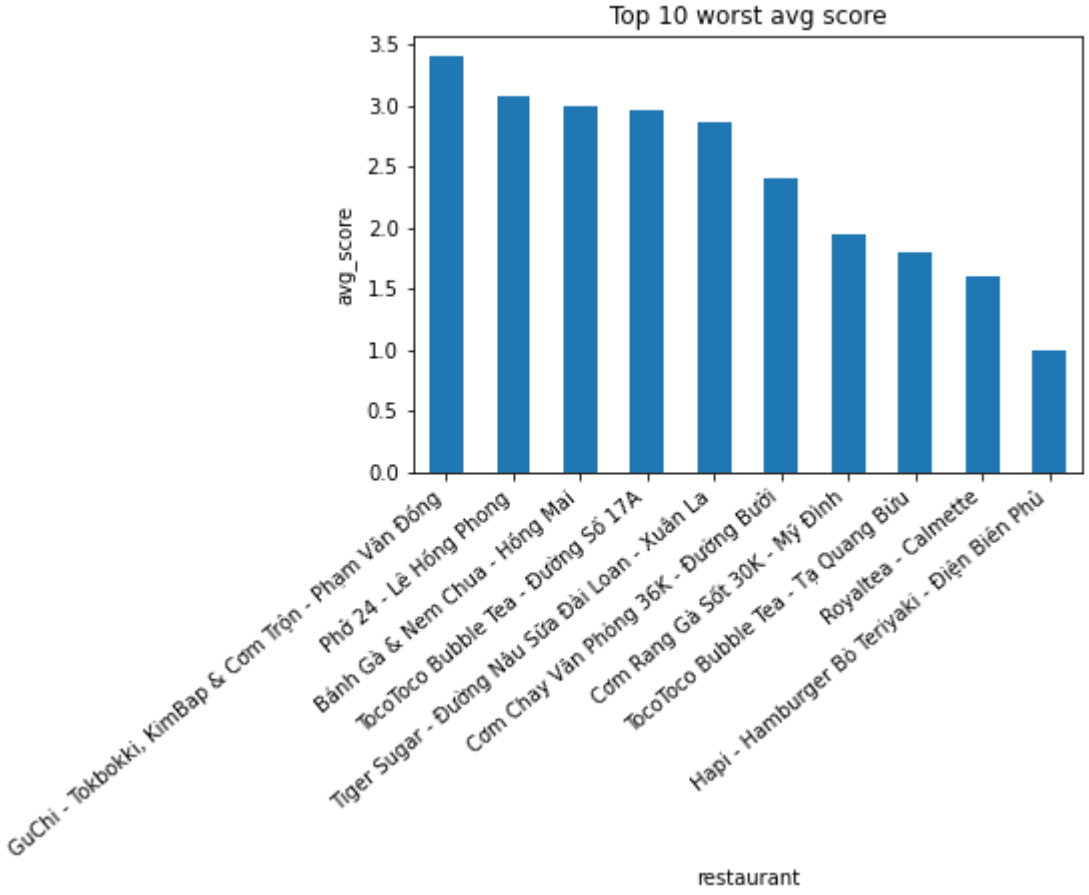
EDA

- Top 10 nhà hàng có điểm trung bình cao nhất:



EDA

- Top 10 nhà hàng có điểm trung bình thấp nhất:



EDA

- Việc đánh giá và cho điểm của khách hàng dựa trên cảm xúc cá nhân và phụ thuộc vào nhiều yếu tố như chất lượng đồ ăn, dịch vụ, thái độ, không gian, giá cả... Có thể trong cùng 1 review có khen điểm này nhưng chưa hài lòng ở điểm khác. Với mỗi khách hàng sẽ có cảm nhận riêng và cách chấm điểm khác nhau.
- Nhóm phân loại reviews của khách hàng theo 2 nhóm:
  - Target 0 (Dislike) với review_score < 6.5
  - Target 1 (Like) với review_score > = 6.5

Nội dung

1. Project Overview
2. EDA
3. **Sentiment Analysis – Logistic Regression Model**
4. Prediction

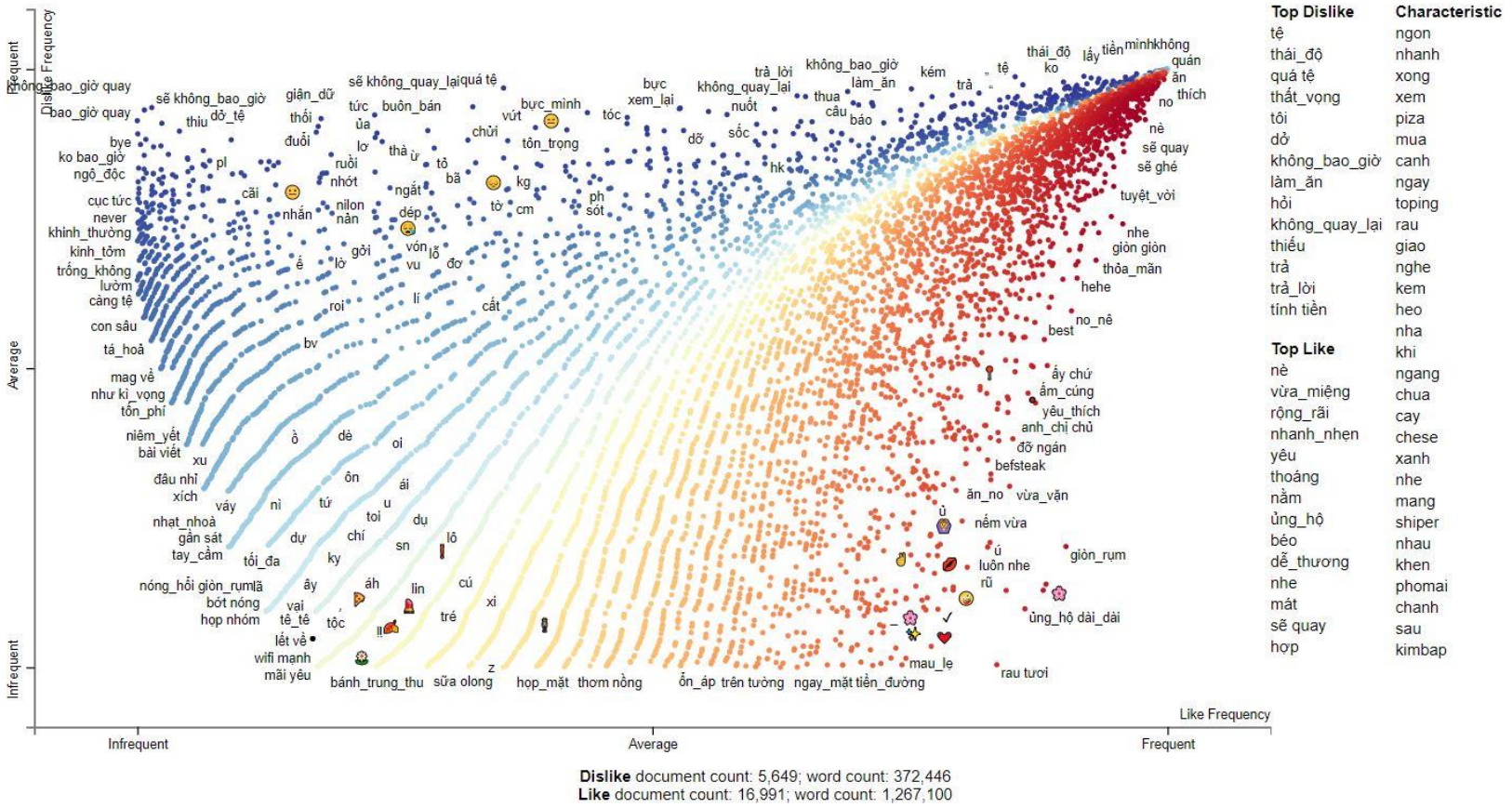
Xử lý dữ liệu

- Xử lý chuyển đổi emoticon, emoji, chữ viết kéo dài đuôi, viết tắt/viết kiểu đặc biệt, dấu câu, loại bỏ các reviews sử dụng tiếng Anh là chính. … tạo column review_text_wt:

	review_text	review_text_wt
0	nhà hàng mới đổi địa chỉ sang hồng hà khá gần ...	nhà_hàng mới đổi địa_chỉ sang hồng hà khá gần ...
1	quán đã chuyển về hồng hà cới phong cách khác ...	quán đã chuyển về hồng hà cới phong_cách khác ...
2	giá niêm yết trên ứng dụng một kiểu giá lúc sh...	giá niêm_yết trên ứng_dụng một kiểu giá lúc sh...
3	xem nhận xét thấy mọi người khen chả cá lã vọn...	xem nhận_xét thấy mọi người khen chả_cá lã vọn...
4	tối nay mới đi ăn quán này món chả cá lã vọng ...	tối nay mới đi ăn quán này món chả_cá lã vọng ...

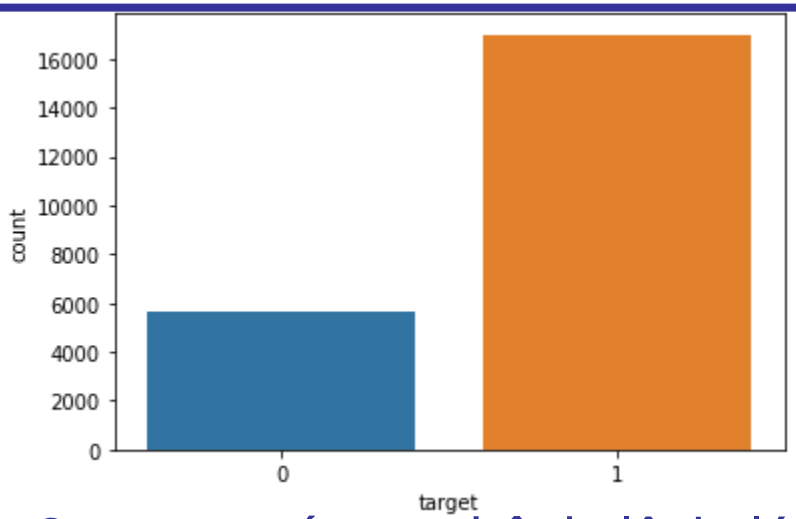
- Dữ liệu còn lại 22640 dòng.

Spacy



- Class Dislike thường xuyên xuất hiện các từ: tệ, thái_độ, quá_tệ, thất_vọng, dở, không_bao_giờ, không_quay_lại,…
- Class Like thường xuyên xuất hiện các từ: vừa_miệng, rộng_rãi, nhanh_nhẹn, yêu, thoáng, ủng_hộ, dễ_thương, …

Sentiment Analysis – Logistic Regression Model



- Dữ liệu giữa 2 target có sự chênh lệch lớn. Có thể cần resample.
- Sử dụng TfidfVectorizer kết hợp stopwords để chuyển text về dạng array với 10000 từ xuất hiện nhiều nhất.
- Sử dụng chi2 để lấy số lượng nhỏ các từ có mối quan hệ đáng kể với biến target, kết quả được 1.218 từ quan trọng nhất.
- Chia dữ liệu thành tập train, test với tỷ lệ 0.7:0.3

Sentiment Analysis – Logistic Regression Model

- Đo hiệu quả của một số mô hình với dữ liệu chưa resample.
- Lựa chọn Logistic Regression (LR) vì có Test Accuracy, F1 Score, AUC-ROC cao nhất. SVC() cũng có Test Accuracy, F1 Score, AUC-ROC xấp xỉ LR nhưng thời gian thực thi quá lâu.

	Model	Test Accuracy	F1 Score	AUC-ROC	Time Fit
0	LogisticRegression	0.882509	0.924274	0.894700	3.681466
1	MultinomialNB	0.863074	0.914679	0.891214	0.137139
2	BernoulliNB	0.851148	0.899253	0.892874	0.419965
3	KNeighborsClassifier	0.750294	0.846571	0.657571	0.047089
4	DecisionTreeClassifier	0.791667	0.862448	0.719120	11.271319
5	RandomForestClassifier	0.863958	0.914666	0.879344	12.842089
6	XGBClassifier	0.853946	0.907307	0.867308	38.084894
7	SVC	0.881920	0.924010	0.893710	737.448229

Sentiment Analysis – Logistic Regression Model

```
lr.score(x_train, y_train)
```

```
0.9006814740030288
```

```
lr.score(x_test, y_test)
```

```
0.8825088339222615
```

```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.84      0.66      0.74      1702
           1       0.89      0.96      0.92      5090

    accuracy                           0.88      6792
   macro avg       0.87      0.81      0.83      6792
weighted avg       0.88      0.88      0.88      6792
```

```
cm = confusion_matrix(y_test, y_pred)
cm
```

```
array([[1124,  578],
       [ 220, 4870]])
```

- LogisticRegression() mặc định và LogisticRegression đã áp dụng tuning parameter đều cho kết quả tương tự nhau.
- Tuy model có accuracy tương đối tốt nhưng kết quả dự đoán cho target 0 (Dislike) chưa chính xác.

Sentiment Analysis – Logistic Regression Model

```
data_train_new.groupby('target').review_text_wt.count()
```

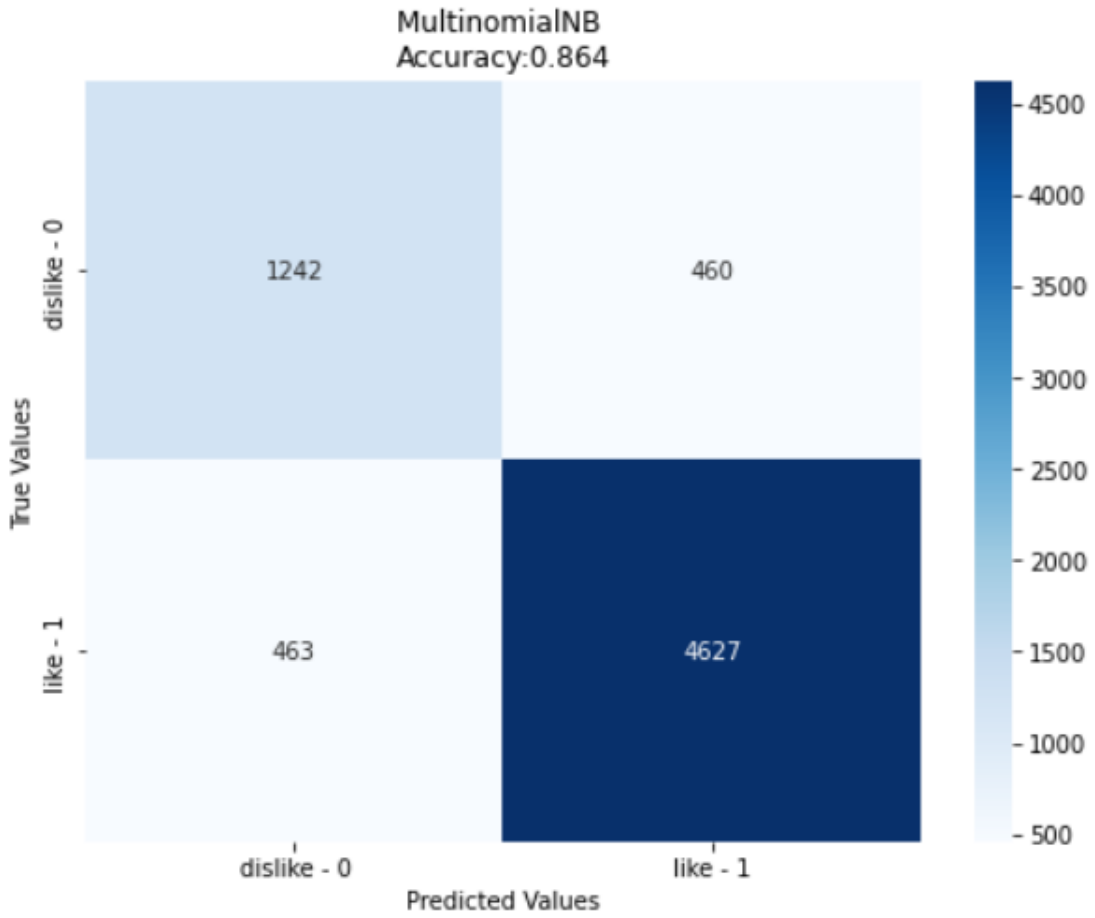
```
target
0    7924
1    7924
```

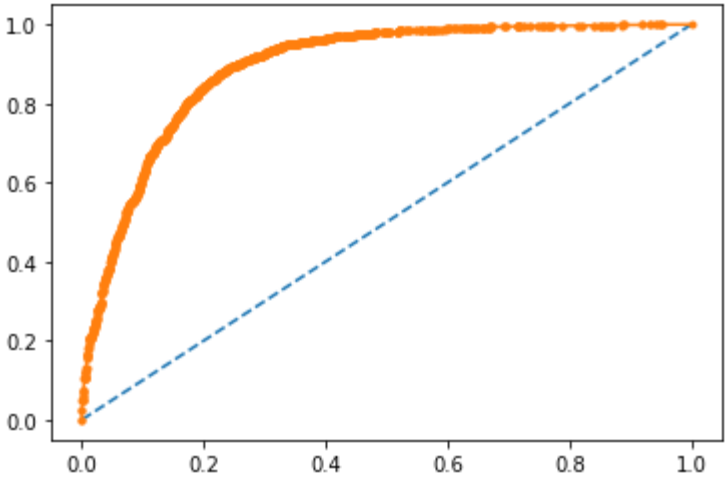
- Đo hiệu quả của một số mô hình với dữ liệu đã resample.
- Dựa vào Test Accuracy, F1 Score, AUC-ROC, Time Fit, em lựa chọn các model sau để build model: LogisticRegression, MultinomialNB, RandomForestClassifier

	Model	Test Accuracy	F1 Score	AUC-ROC	Time Fit
0	LogisticRegression	0.845701	0.893474	0.893302	3.364313
1	MultinomialNB	0.864105	0.909305	0.892086	0.095594
2	BernoulliNB	0.820966	0.874146	0.891726	0.230634
3	KNeighborsClassifier	0.648999	0.751511	0.637744	0.021819
4	DecisionTreeClassifier	0.749411	0.825257	0.717317	8.807775
5	RandomForestClassifier	0.857479	0.906564	0.878848	13.434146
6	XGBClassifier	0.799176	0.859207	0.865653	38.659609

Sentiment Analysis – Logistic Regression Model

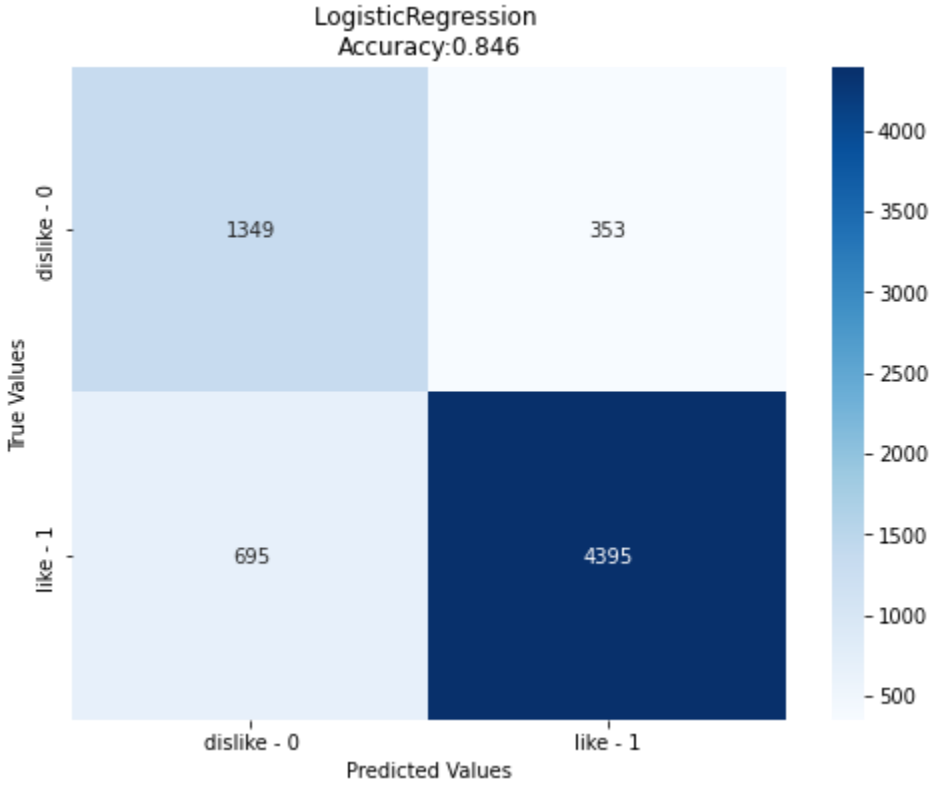
- MultinomialNB: Model dự đoán khá tốt, accuracy tương đối cao.

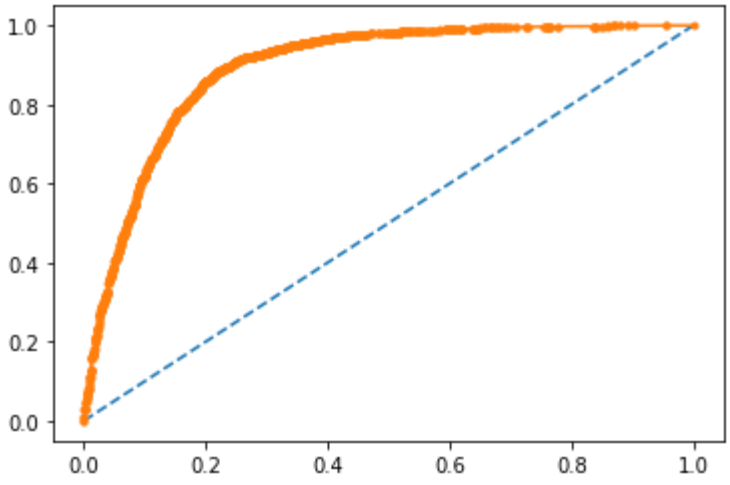




Sentiment Analysis – Logistic Regression Model

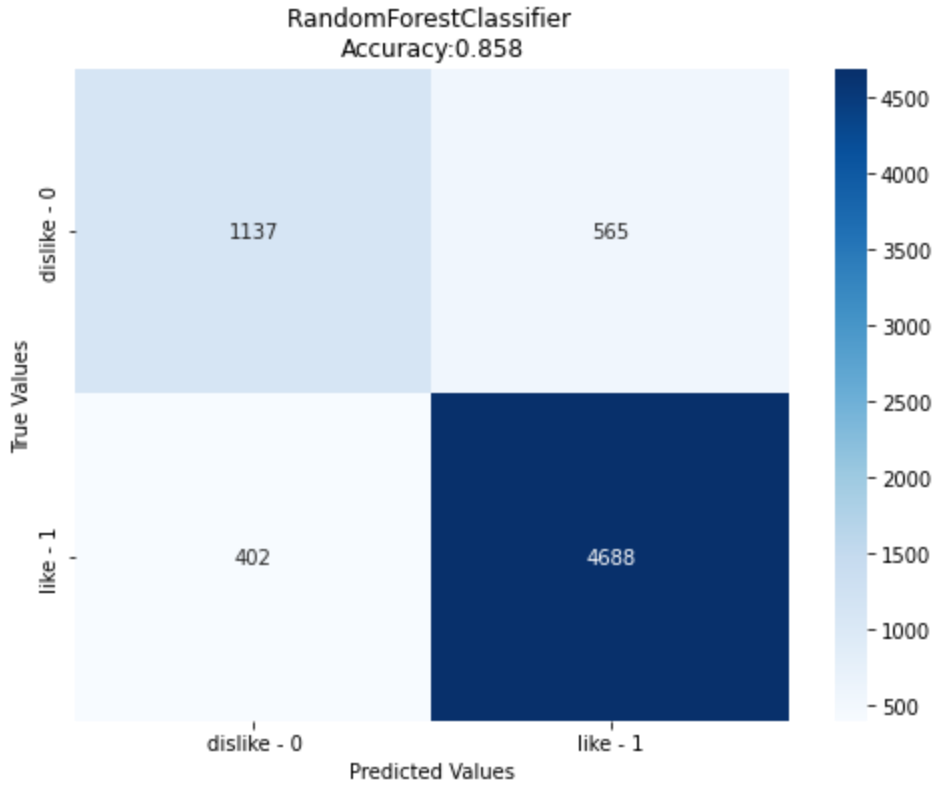
- LogisticRegression: Model dự đoán khá tốt ở cả 2 target, accuracy thấp hơn MultinomialNB một chút.

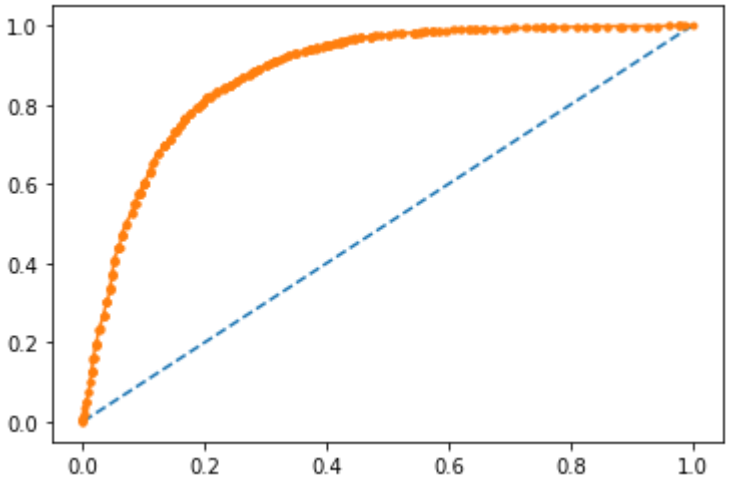




Sentiment Analysis – Logistic Regression Model

- RandomForestClassifier: Model dự đoán khá tốt, accuracy tương đối cao.





Sentiment Analysis – Logistic Regression Model

- Thử tuning parameter 3 thuật toán này nhưng kết quả không tốt hơn so với model gốc.
- Lựa chọn model LogisticRegression để dự đoán vì có dự đoán đồng đều ở cả 2 class, tuy accuracy có thấp hơn một chút so với 2 thuật toán còn lại.

Nội dung

1. Project Overview
2. EDA
3. Sentiment Analysis – Logistic Regression Model
4. Prediction

Prediction

```
[ ] x_new = 'Đặt đơn bôm đợt 12/12 nên được tặng cái hold cuo khá cute của ShopeeFood. Trà Phúc Long thì mê xưa giờ rồi, trà đậm vị uống chất lượng uốn
```

```
[ ] x_new = xu_ly(x_new)
```

```
[ ] x_new = count.transform(np.array([x_new])).toarray()
    x_new
```

```
array([[0., 0., 0., ..., 0., 0., 0.]])
```

```
y_hat = my_model.predict(x_new)
y_hat
```

```
array([1])
```

- y dự đoán là loại 1 (Like), thực tế, người đánh giá cho 8.8đ