

BÁO CÁO ĐỒ ÁN CUỐI KHÓA KHOA HỌC DỮ LIỆU – DSP305

DỰ ĐOÁN LƯỢNG KHÁCH DU LỊCH

Học viên: PHẠM CÔNG SỰ – FX16803

Mục lục

I. Phân tích dữ liệu khám phá

1. Hiểu biết về nghiệp vụ
2. Hiểu biết về dữ liệu
3. Phân tích dữ liệu
4. Phương pháp giải quyết vấn đề

II. Lập mô hình và đánh giá

1. Cài đặt tham số
2. Chia dữ liệu huấn luyện, đánh giá
3. Đánh giá mô hình

III. Cải thiện mô hình

1. Tạo các đặc trưng mới
2. Tinh chỉnh các tham số mô hình
3. So sánh các kết quả

IV. Kết luận

I

Phân tích dữ liệu khám phá



Hiểu biết về nghiệp vụ

Vấn đề của đồ án:

- Du lịch là một ngành kinh tế quan trọng, đóng góp đáng kể cho nền kinh tế của nhiều quốc gia. Tuy nhiên lượng khách du lịch luôn thay đổi và có thể bị ảnh hưởng bởi nhiều yếu tố. Dự báo nhu cầu du lịch chính xác sẽ giúp các doanh nghiệp du lịch và khách sạn giảm thiểu những chi phí không cần thiết và tận dụng được tối đa các cơ hội kinh doanh.
- Đồ án được lấy từ một cuộc thi trên kaggle, yêu cầu xây dựng mô hình dự đoán lượng khách du lịch để cung cấp thông tin chính xác về nhu cầu du lịch trong tương lai của Nhật Bản, giúp các doanh nghiệp và cơ quan liên quan đưa ra các quyết định kinh doanh và phát triển phù hợp, góp phần thúc đẩy sự phát triển của ngành du lịch.
- Chi tiết: <https://www.kaggle.com/competitions/prediction-of-tourist-arrivals/overview>



Hiểu biết về nghiệp vụ

Mục tiêu:

- Dự đoán lượng khách du lịch sẽ tới trong vòng 1 tháng, từ ngày 01/07/2019 tới ngày 31/07/2019.

Các yếu tố có thể ảnh hưởng đến lượng khách du lịch:

- Thời gian
- Vị trí địa lý: xa/gần, vùng miền...
- Đặc điểm của nơi đến: đông dân/thưa dân, thành phố lớn/nông thôn...
- Sự kiện, lễ hội
- Thời tiết
- Các yếu tố khác



Hiểu biết về nghiệp vụ

Tiêu chí đánh giá:

- Đồ án được đánh giá bằng: **RMSE** (theo quy định của cuộc thi trên kaggle)
- RMSE là viết tắt của Root Mean Square Error (Sai số bình phương trung bình gốc). Đây là một thước đo thống kê được sử dụng để đánh giá mức độ chính xác của một mô hình dự đoán so với dữ liệu thực tế.
- Ý nghĩa của RMSE:
 - RMSE càng nhỏ thì mô hình dự đoán càng chính xác.
 - RMSE càng lớn thì mô hình dự đoán càng kém chính xác.
- Đồ án được đánh giá là đạt yêu cầu khi RMSE nhỏ hơn hoặc bằng kết quả **top 20** trên bảng xếp hạng của cuộc thi trên kaggle vào thời điểm bắt đầu làm đồ án (**RMSE <= 78.33**)



Hiểu biết về dữ liệu

Dữ liệu huấn luyện (train_df.csv):

- Bảng chứa dữ liệu bao gồm 12 biến giải thích và 1 biến mục tiêu (số lượng khách du lịch đã đến).
- Số lượng dữ liệu huấn luyện: **132,192**

	Trường (fields)	Định nghĩa
1	id	ID của bản ghi
2	date	Năm, tháng, ngày
3	tourist_area	Vùng du lịch
4	spot_facility	Điểm cơ sở
5	area	Loại khu vực
6	city	Loại thành phố
7	type	type
8	category	category
9	tourism_index	Chỉ số du lịch
10	info	Info
11	event	Event
12	weather_index	Chỉ số thời tiết
13	tourist_arrivals	Số lượng khách du lịch (biến mục tiêu)

Hiểu biết về dữ liệu

Dữ liệu kiểm tra (test_df.csv):

- Tương tự bảng dữ liệu huấn luyện nhưng không có “tourist_arrivals”.
- Số lượng dữ liệu kiểm tra: **13,392**

	Trường (fields)	Định nghĩa
1	id	ID của bản ghi
2	date	Năm, tháng, ngày
3	tourist_area	Vùng du lịch
4	spot_facility	Điểm cơ sở
5	area	Loại khu vực
6	city	Loại thành phố
7	type	type
8	category	category
9	tourism_index	Chỉ số du lịch
10	info	Info
11	event	Event
12	weather_index	Chỉ số thời tiết

Phân tích dữ liệu

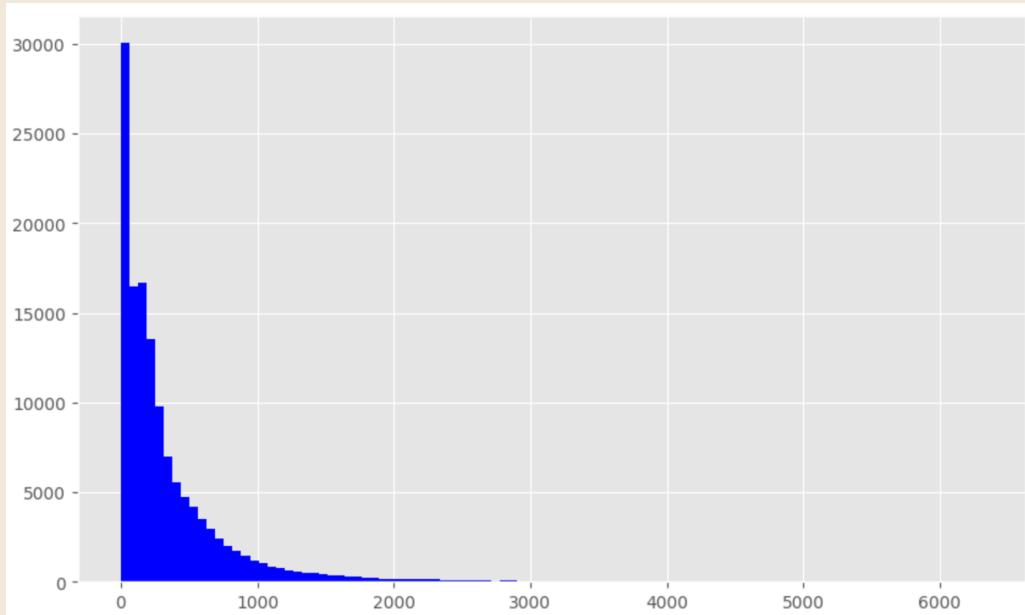
Mô tả dữ liệu huấn luyện (train_df.csv):

		dtypes	missing#	missing%	uniques	count
	id	int64	0	0.000000	132192	132192
	date	datetime64[ns]	0	0.000000	303	132192
	tourist_area	int64	0	0.000000	54	132192
	spot_facility	object	0	0.000000	8	132192
	tourist_arrivals	int64	0	0.000000	3071	132192
	area	object	0	0.000000	22	132192
	city	object	0	0.000000	16	132192
	type	object	0	0.000000	5	132192
	category	int64	0	0.000000	17	132192
	tourism_index	float64	3992	0.030198	3609	128200
	info	object	0	0.000000	4	132192
	event	object	0	0.000000	7	132192
	weather_index	float64	41040	0.310458	190	91152

Phân tích dữ liệu

Mô tả dữ liệu huấn luyện (train_df.csv):

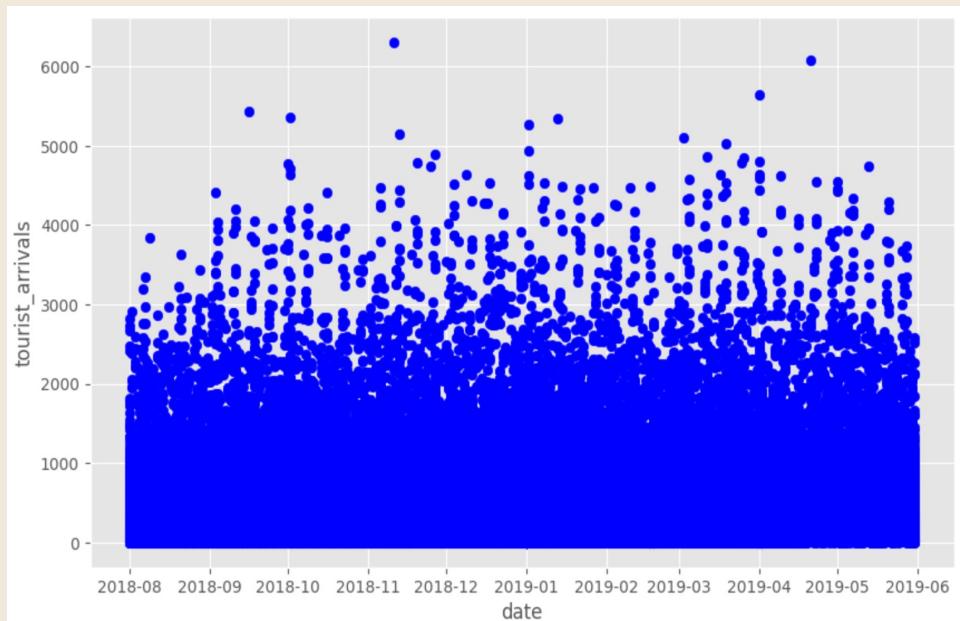
- tourist_arrivals (output)
- Nhận xét: phần lớn giá trị tập trung trong khoảng 0-1000



Phân tích dữ liệu

Mô tả dữ liệu huấn luyện (train_df.csv):

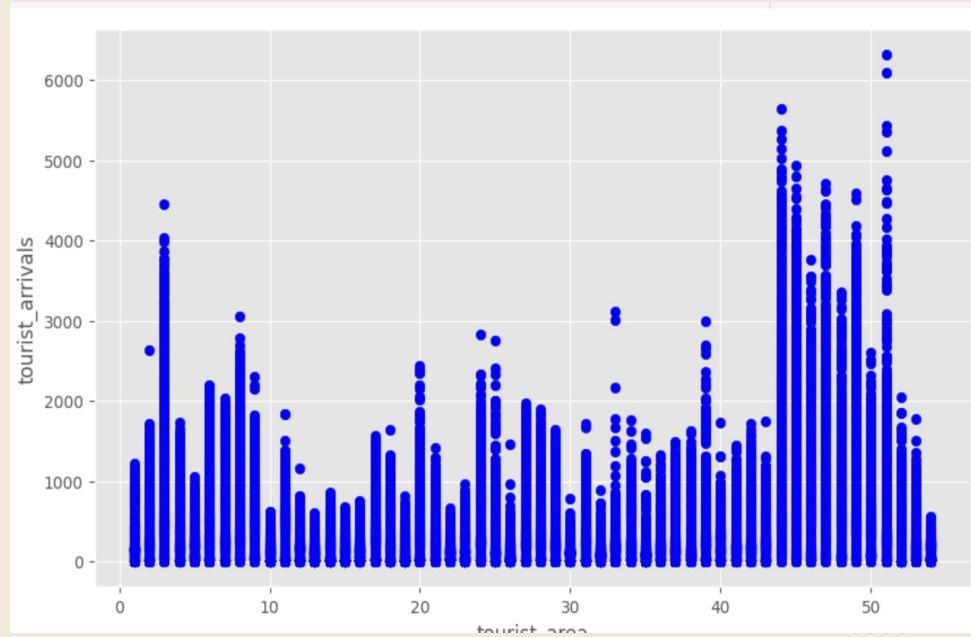
- date
- Nhận xét:
 - Đa số các ngày đều có lượng khách du lịch từ 3000 trở xuống.
 - Ngày có lượng khách du lịch cao nhất là khoảng 6200.
 - tourist_arrivals có thể phụ thuộc vào date



Phân tích dữ liệu

Mô tả dữ liệu huấn luyện (train_df.csv):

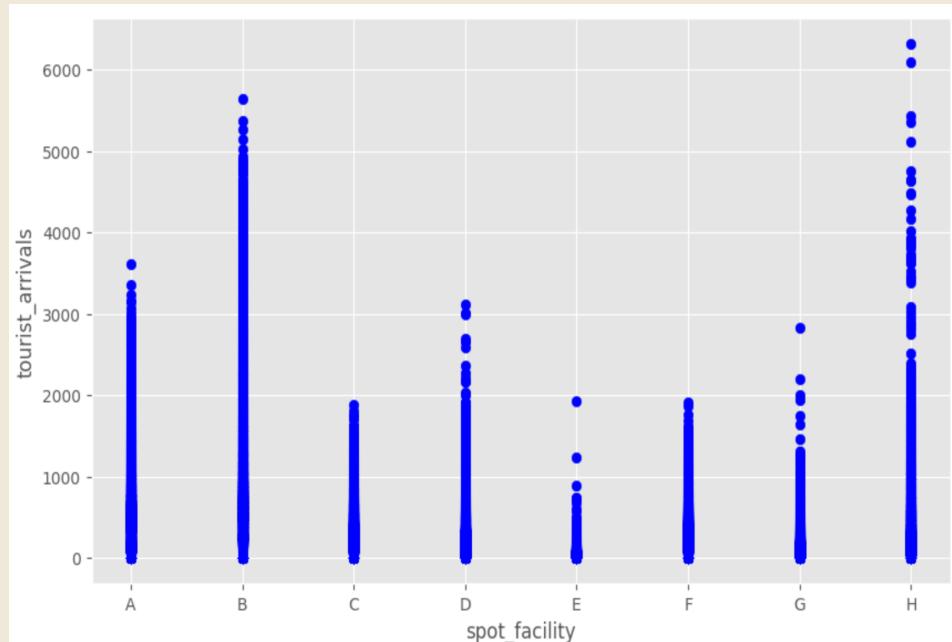
- tourist_area
- Nhận xét:
 - Có tổng cộng 54 loại tourist_area khác nhau
 - Khách du lịch có xu hướng tới tourist_area từ 44 tới 54
 - tourist_arrivals có thể phụ thuộc vào tourist_area



Phân tích dữ liệu

Mô tả dữ liệu huấn luyện (train_df.csv):

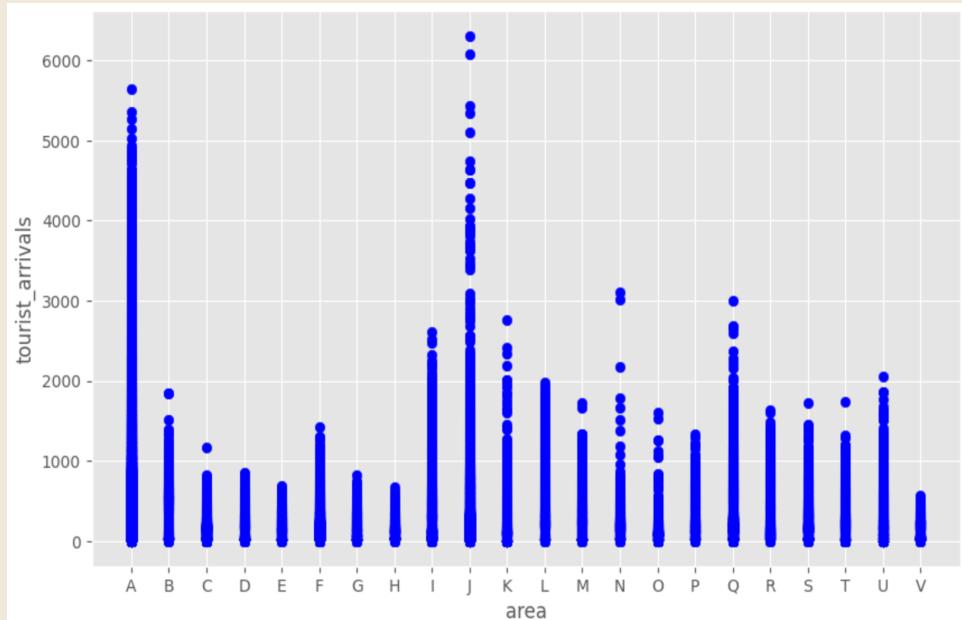
- `spot_facility`
- Nhận xét:
 - Có 8 loại `spot_facility` khác nhau (A, B, C, D, E, F, G, H)
 - `tourist_arrivals` có xu hướng cao ở `spot_facility` B và H, có xu hướng thấp ở `spot_facility` E
 - `tourist_arrivals` có thể phụ thuộc vào `spot_facility`



Phân tích dữ liệu

Mô tả dữ liệu huấn luyện (train_df.csv):

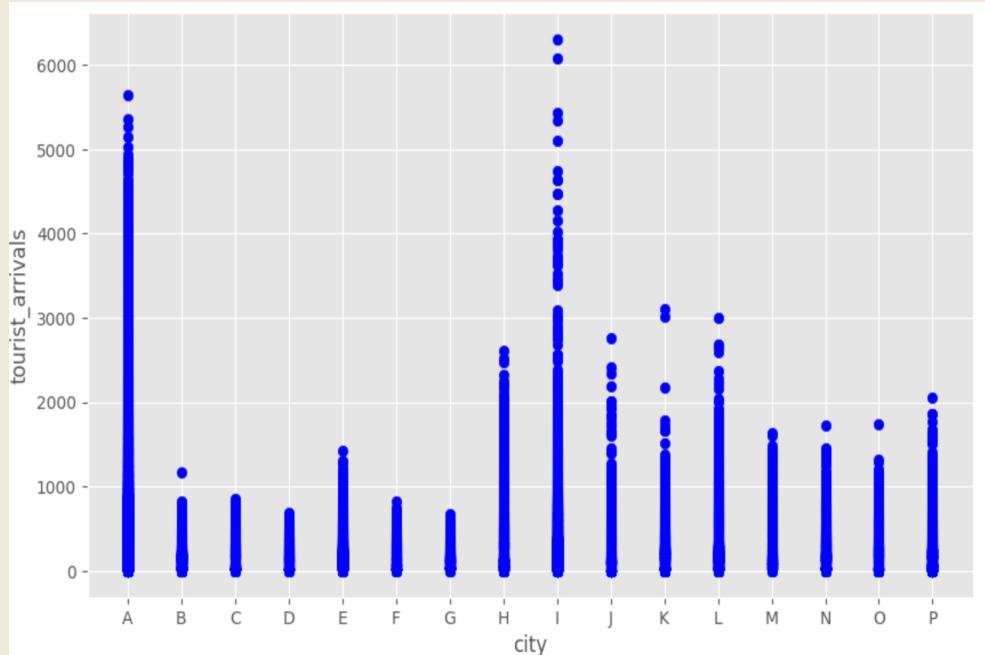
- area
- Nhận xét:
 - Có tổng cộng 22 loại area khác nhau
 - tourist_arrivals có xu hướng cao ở area A và J, trong khi đó có xu hướng thấp ở C, D, E, G, H, V
 - tourist_arrivals có thể phụ thuộc vào area



Phân tích dữ liệu

Mô tả dữ liệu huấn luyện (train_df.csv):

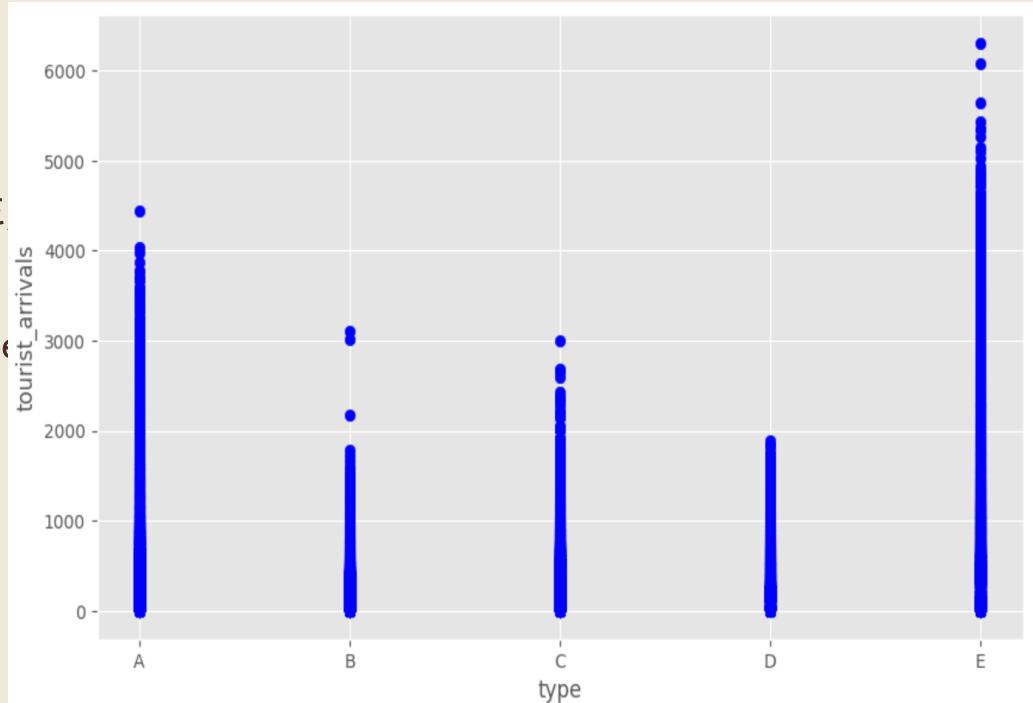
- city
- Nhận xét:
 - Có 16 loại city khác nhau
 - tourist_arrivals có xu hướng cao ở city A và I, trong khi đó có xu hướng thấp ở B, C, D, F, G
 - tourist_arrivals có thể phụ thuộc vào city



Phân tích dữ liệu

Mô tả dữ liệu huấn luyện (train_df.csv):

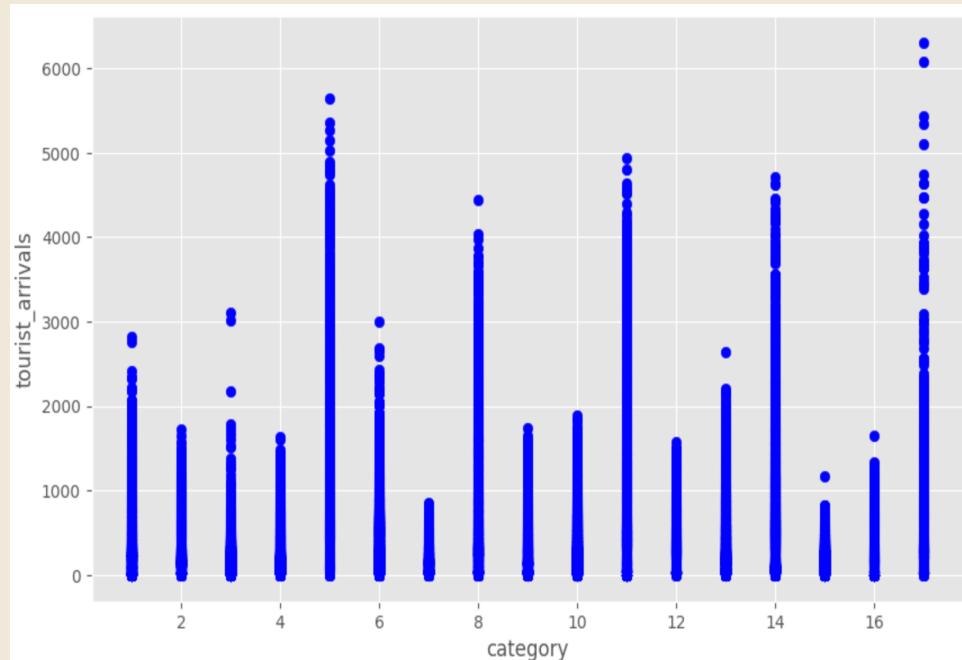
- type
- Nhận xét:
 - Có 5 loại type khác nhau
 - tourist_arrivals có xu hướng cao ở type E
trong khi thấp ở type D
 - tourist_arrivals có thể phụ thuộc vào type



Phân tích dữ liệu

Mô tả dữ liệu huấn luyện (train_df.csv):

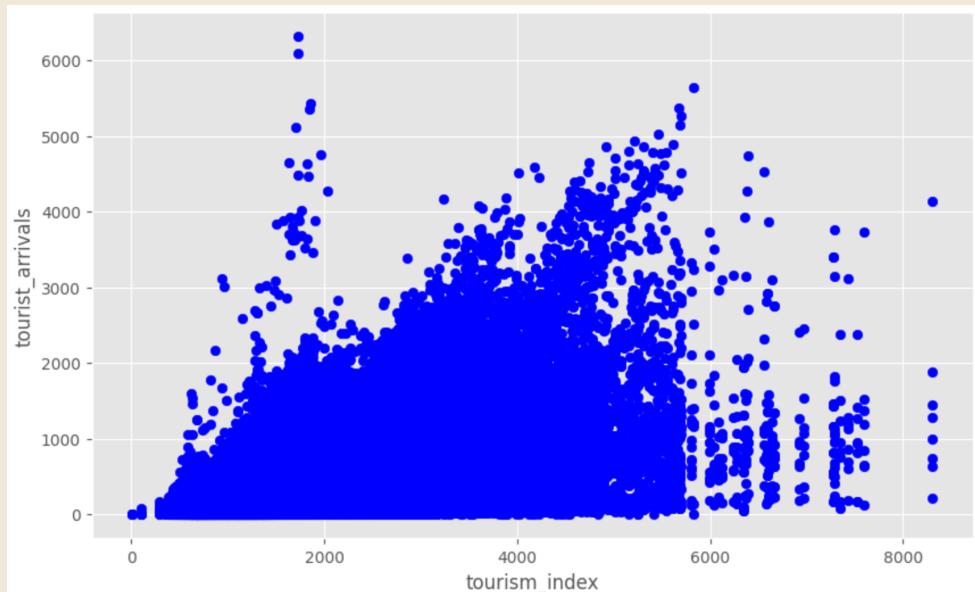
- category
- Nhận xét:
 - Có tổng cộng 17 loại category khác nhau
 - tourist_arivals có xu hướng cao ở category 3, 11, 14, 17, trong khi đó có xu hướng thấp ở category 7, 15
 - tourist_arrivals có thể phụ thuộc vào category



Phân tích dữ liệu

Mô tả dữ liệu huấn luyện (train_df.csv):

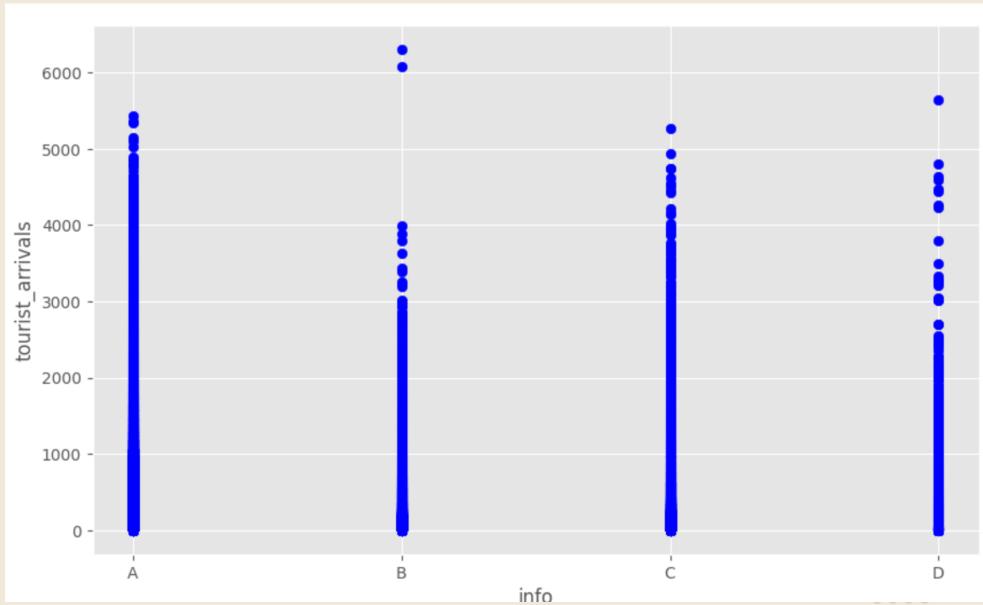
- **tourism_index**
- Nhận xét:
 - tourist_arrivals có xu hướng tập trung ở tourism_index dưới 6000
 - tourist_arrivals có thể phụ thuộc vào tourism_index



Phân tích dữ liệu

Mô tả dữ liệu huấn luyện (train_df.csv):

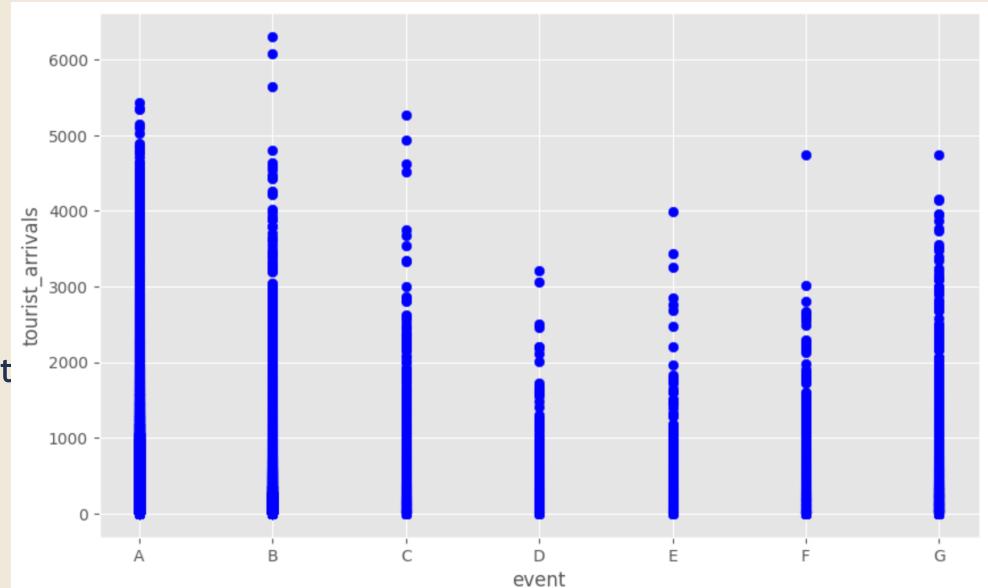
- info
- Nhận xét:
 - Có 4 loại info khác nhau
 - tourist_arrivals có xu hướng cao ở A và C, trong khi thấp hơn ở B và D
 - tourist_arrivals có thể phụ thuộc vào info



Phân tích dữ liệu

Mô tả dữ liệu huấn luyện (train_df.csv):

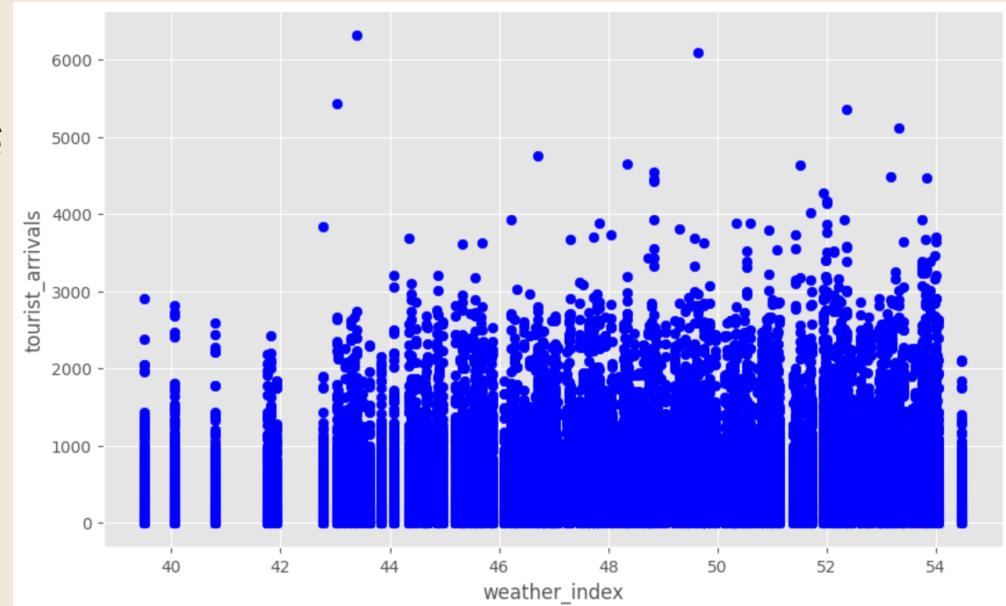
- event
- Nhận xét:
 - Có tổng cộng 7 loại event khác nhau
 - tourist_arrivals có xu hướng cao ở A và B, trong khi thấp ở D
 - tourist_arrivals có thể phụ thuộc vào event



Phân tích dữ liệu

Mô tả dữ liệu huấn luyện (train_df.csv):

- **weather_index**
- Nhận xét:
 - Tourist_arrivals có xu hướng tập trung về weather_index từ khoảng 43 tới 54
 - tourist_arrivals có thể phụ thuộc vào weather_index



Phân tích dữ liệu

Mô tả dữ liệu kiểm tra (test_df.csv):

		dtypes	missing#	missing%	uniques	count
	id	int64	0	0.000000	13392	13392
	date	datetime64[ns]	0	0.000000	31	13392
	tourist_area	int64	0	0.000000	54	13392
	spot_facility	object	0	0.000000	8	13392
	area	object	0	0.000000	22	13392
	city	object	0	0.000000	16	13392
	type	object	0	0.000000	5	13392
	category	int64	0	0.000000	17	13392
	tourism_index	int64	0	0.000000	1216	13392
	info	object	0	0.000000	2	13392
	event	object	0	0.000000	3	13392
	weather_index	float64	5184	0.387097	19	8208

Phân tích dữ liệu

Biến đổi dữ liệu:

- Tạo thêm đặc trưng “year”, “month”, “day” từ đặc trưng ban đầu là “date”
- Sử dụng Label Encoding để chuyển đổi các dữ liệu dạng text thành dạng số cho các đặc trưng: 'spot_facility', 'area', 'city', 'type', 'info', 'event'



Phân tích dữ liệu

Biến đổi dữ liệu:

train_df

		dtypes	missing#	missing%	uniques	count
	id	int64	0	0.000000	132192	132192
	date	datetime64[ns]	0	0.000000	303	132192
	tourist_area	int64	0	0.000000	54	132192
	spot_facility	int64	0	0.000000	8	132192
	tourist_arrivals	int64	0	0.000000	3071	132192
	area	int64	0	0.000000	22	132192
	city	int64	0	0.000000	16	132192
	type	int64	0	0.000000	5	132192
	category	int64	0	0.000000	17	132192
	tourism_index	float64	3992	0.030198	3609	128200
	info	int64	0	0.000000	4	132192
	event	int64	0	0.000000	7	132192
	weather_index	float64	41040	0.310458	190	91152
	year	int64	0	0.000000	2	132192
	month	int64	0	0.000000	10	132192
	day	int64	0	0.000000	31	132192

test_df

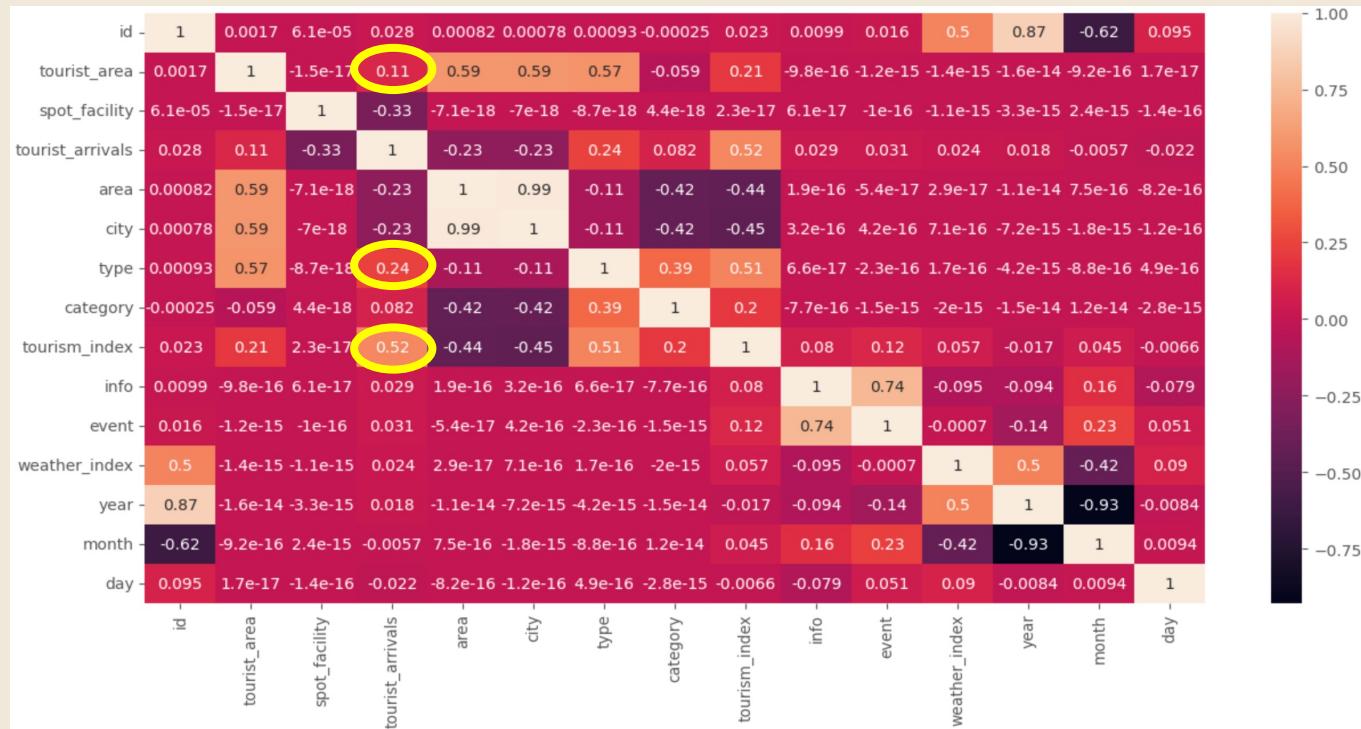
		dtypes	missing#	missing%	uniques	count
	id	int64	0	0.000000	13392	13392
	date	datetime64[ns]	0	0.000000	31	13392
	tourist_area	int64	0	0.000000	54	13392
	spot_facility	int64	0	0.000000	8	13392
	area	int64	0	0.000000	22	13392
	city	int64	0	0.000000	16	13392
	type	int64	0	0.000000	5	13392
	category	int64	0	0.000000	17	13392
	tourism_index	int64	0	0.000000	1216	13392
	info	int64	0	0.000000	2	13392
	event	int64	0	0.000000	3	13392
	weather_index	float64	5184	0.387097	19	8208
	year	int64	0	0.000000	1	13392
	month	int64	0	0.000000	1	13392
	day	int64	0	0.000000	31	13392

Phân tích dữ liệu

Tương quan giữa các đặc trưng:

- Với biến phụ thuộc (output) là “tourist_arrivals”, một số đặc trưng (feature) có hệ số tương quan cao nhất là:

- tourism_index
- type
- tourist_area



Phân tích dữ liệu

Lựa chọn đặc trưng:

- Loại bỏ “weather_index” khỏi dữ liệu vì có hơn 30% giá trị bị thiếu ở cả tập dữ liệu huấn luyện và kiểm tra
- Loại bỏ “id”, “date”
- Loại bỏ các dòng có giá trị bị thiếu của “tourism_index”, giữ lại trường này vì tỷ lệ giá trị bị thiếu thấp, và có hệ số tương quan cao với “tourist_arivals”

	Feature	Correlation (absolute value)
1	id	0.027750
2	tourist_area	0.112900
3	spot_facility	0.325611
4	area	0.227062
5	city	0.229929
6	type	0.244973
7	category	0.082123
8	tourism_index	0.519752
9	info	0.029452
10	event	0.030675
11	weather_index	0.024104
12	year	0.017656
13	month	0.005725
14	day	0.022258

Phương án giải quyết vấn đề

Thuật toán:

- Áp dụng thuật toán học máy Gradient Boosting, thử nghiệm trên 3 framework Gradient Boosting phổ biến:
 - LightGBM (<https://lightgbm.readthedocs.io/en/stable/>)
 - XGBoost (<https://xgboost.readthedocs.io/en/stable/>)
 - CatBoost (<https://catboost.ai/>)



Phương án giải quyết vấn đề

Kỹ thuật:

- Ngôn ngữ lập trình: Python
- Library/Package/Framework:
 - Xử lý dữ liệu: pandas, numpy, datetime, jpholiday
 - Thống kê và phân tích dữ liệu: numpy, sklearn
 - Học máy: sklearn, keras, optuna, lightgbm, xgboost, catboost
 - Vẽ đồ thị: matplotlib, seaborn



II

Lập mô hình và đánh giá



Cài đặt tham số

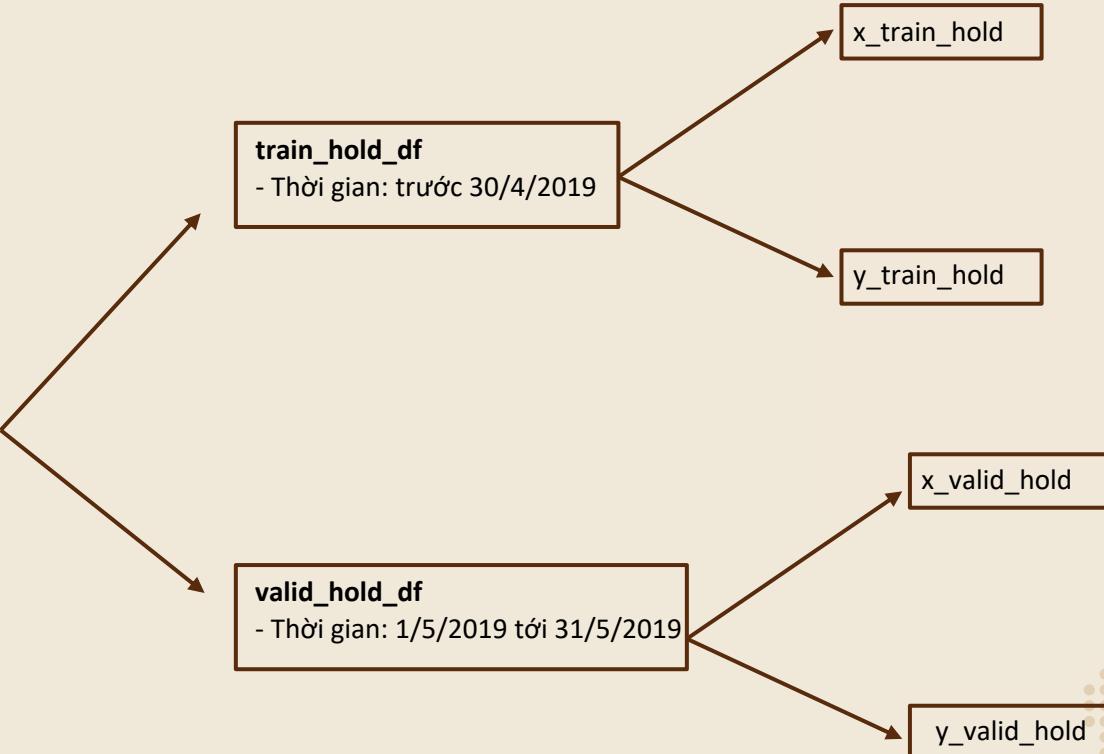
- Các tham số sẽ tiến hành cài đặt và điều chỉnh:

Định nghĩa	Tên tham số		
	LightGBM	XGBoost	CatBoost
Tỷ lệ phần trăm hàng được sử dụng cho mỗi lần lặp lại xây dựng cây	bagging_fraction	subsample	subsample
Tốc độ học tập của mô hình	learning_rate	learning_rate	learning_rate
Số cây	num_iterations	n_estimators	iterations
Độ sâu tối đa của cây	max_depth	max_depth	depth

Chia dữ liệu huấn luyện, kiểm định

Feature	Output
tourist_area	
spot_facility	
area	
city	
type	
category	
tourism_index	
info	
event	
year	
month	
day	

train_df



Đánh giá mô hình

Phương pháp đánh giá:

- Mô hình không được quá khớp trên tập huấn luyện so với tập kiểm định
- Mô hình có RMSE trên tập kiểm định nhỏ nhất là mô hình tốt nhất
- Mô hình có RMSE nhỏ hơn tiêu chí đánh giá là mô hình đạt yêu cầu ($\text{RMSE} \leq 78.33$)

Huấn luyện mô hình:

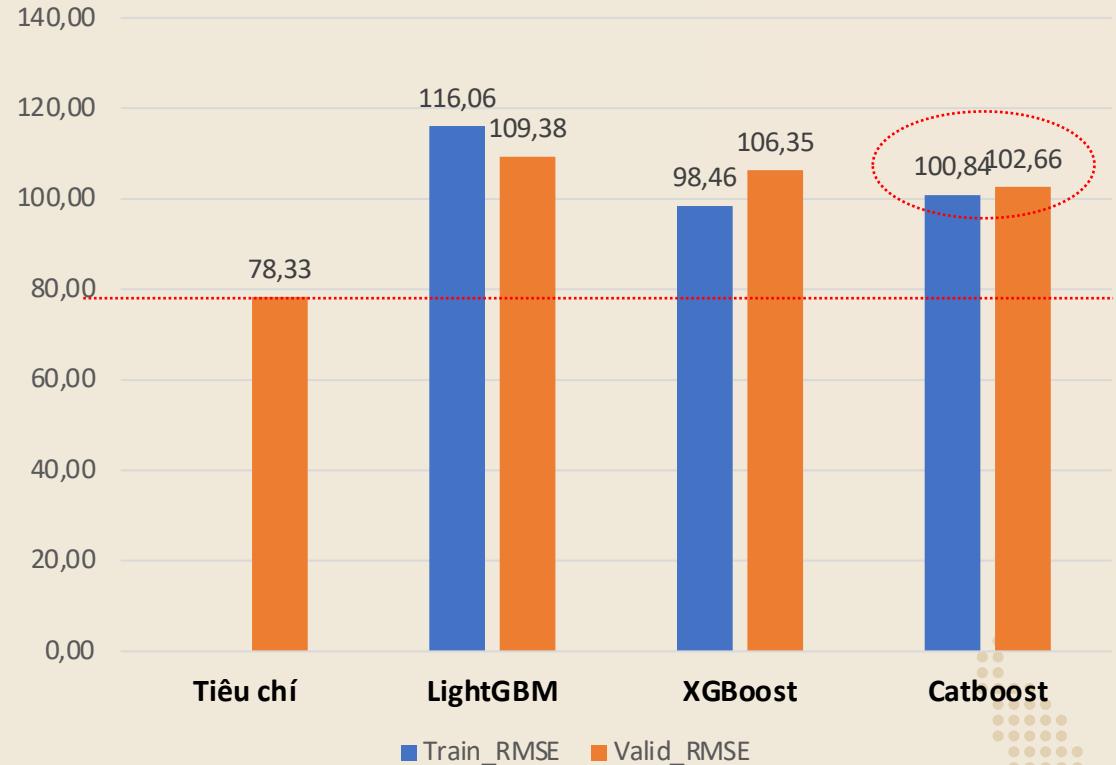
- Huấn luyện mô hình LightGBM, XGBoost và CatBoost với các giá trị tham số mặc định:
 - subsample
 - learning_rate
 - num_iterations
 - max_depth
- `early_stopping_rounds = 100`



Đánh giá mô hình

Kết quả đánh giá:

- Chưa đạt tiêu chí RMSE ≤ 78.33
- Các mô hình không bị overfitting quá nhiều
- Kết quả trên tập kiểm định tốt nhất:
 $\text{CatBoost} = 102.66$
- Kết quả trên tập kiểm định kém nhất: $\text{LightGBM} = 109.38$



III

Cải thiện mô hình



Tạo các đặc trưng mới

Tạo thêm các đặc trưng cho biến mục tiêu

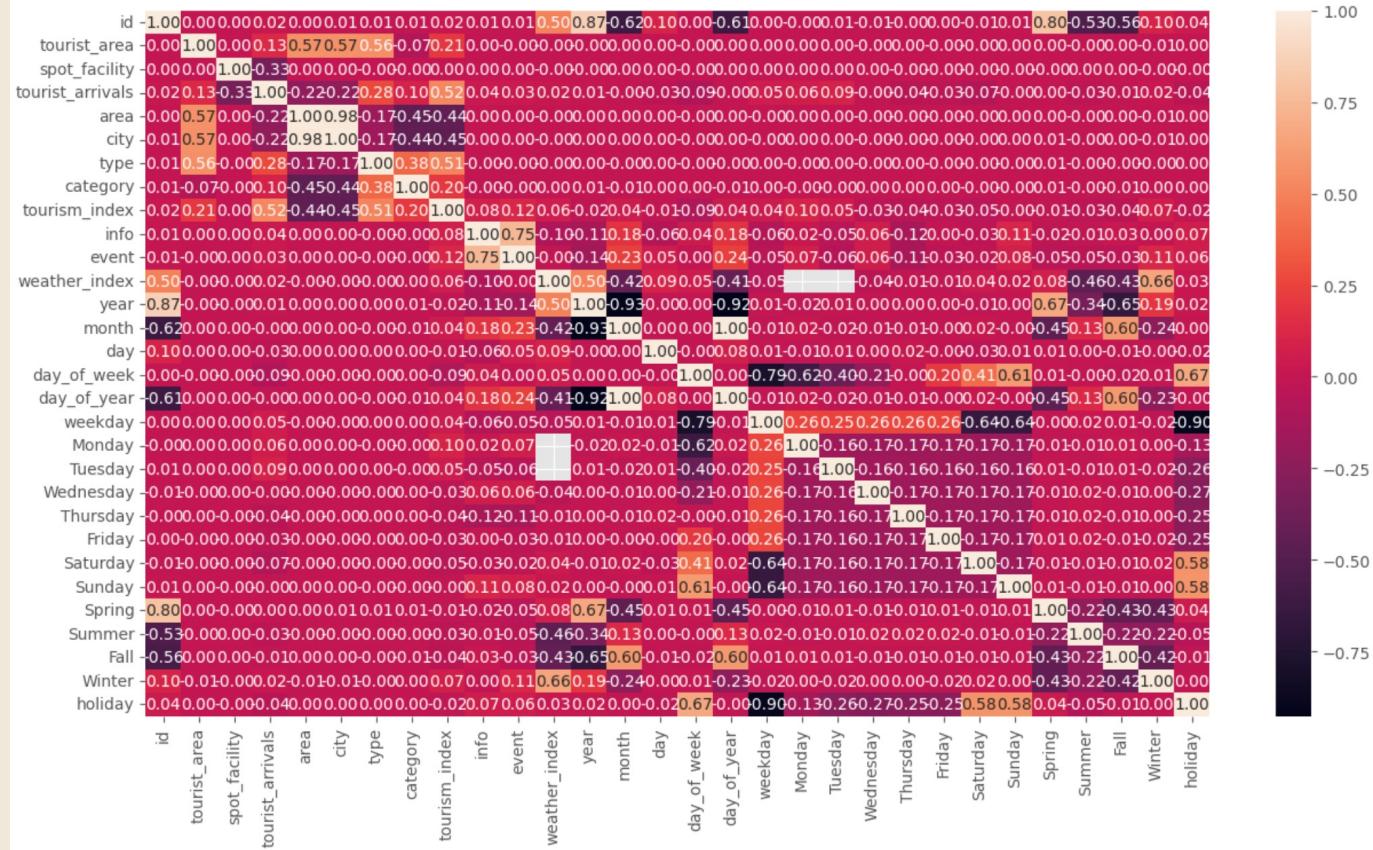
- Xây dựng các đặc trưng mới từ đặc trưng ban đầu nhằm tăng mức độ giải thích biến mục tiêu:

STT	Feature	Định nghĩa
1	day_of_week	ngày thứ bao nhiêu trong tuần
2	day_of_year	ngày thứ bao nhiêu trong năm
3	Weekday	có phải ngày trong tuần không (từ thứ 2 tới thứ 5)
4	Monday	có phải thứ 2 không
5	Tuesday	có phải thứ 3 không
6	Wednesday	có phải thứ 4 không
7	Thursday	có phải thứ 5 không
8	Friday	có phải thứ 6 không
9	Saturday	có phải thứ 7 không
10	Sunday	có phải chủ nhật không
11	Spring	có phải mùa xuân không (tháng 3, 4, 5)
12	Summer	có phải mùa hè không (tháng 6, 7, 8)
13	Fall	có phải mùa thu không (tháng 9, 10, 11)
14	Winter	có phải mùa đông không (tháng 12, 1, 2)
15	Holiday	có phải ngày nghỉ lễ không (các ngày nghỉ lễ của Nhật Bản, thứ 7, chủ nhật)

Tạo các đặc trưng mới

Lựa chọn đặc trưng

- 15 đặc trưng mới có tương quan tốt với các đặc trưng ban đầu
- Bổ sung 15 đặc trưng mới vào 12 đặc trưng ban đầu



Tạo các đặc trưng mới

Huấn luyện mô hình:

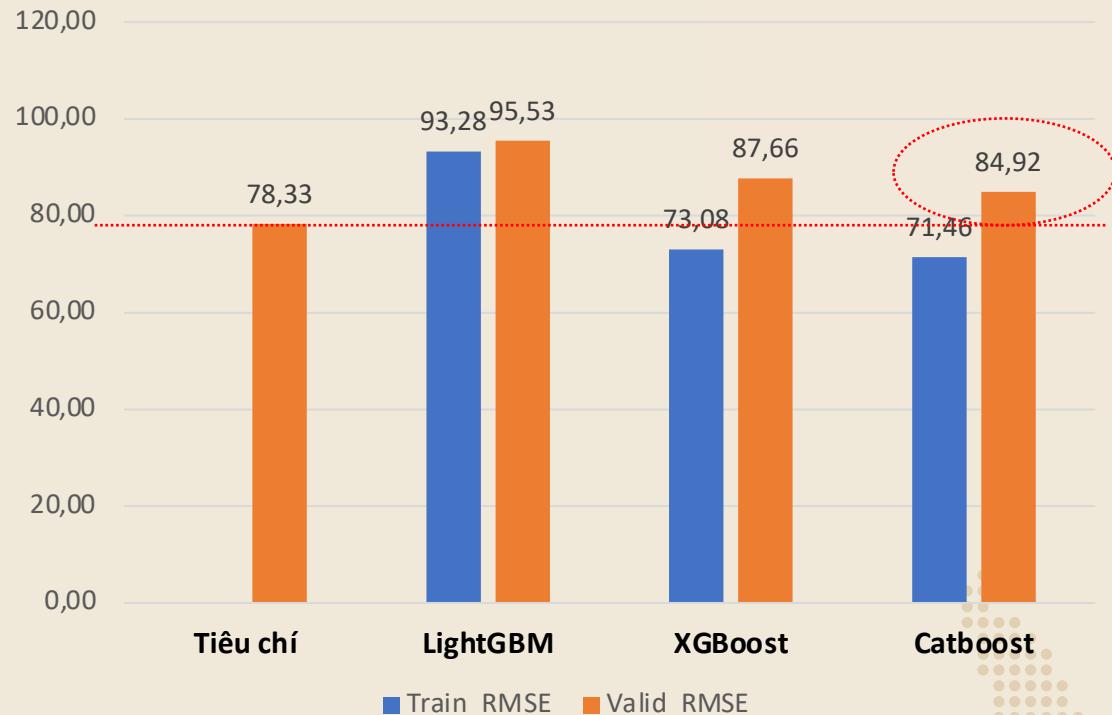
- Huấn luyện lại mô hình LightGBM, XGBoost và CatBoost với các giá trị tham số mặc định:
 - subsample
 - learning_rate
 - num_iterations
 - max_depth
- early_stopping_rounds = 100



Tạo các đặc trưng mới

Kết quả

- Chưa mô hình nào đạt tiêu chí
RMSE <= 78.83
- Tuy nhiên đã cải thiện so với các mô hình ban đầu:
 - Kết quả trên tập kiểm định tốt nhất: CatBoost = 84.92 (giảm 17.74)
 - Kết quả trên tập kiểm định kém nhất: LightGBM = 95.53 (giảm 13.85)



Tinh chỉnh các tham số

Chọn tham số tinh chỉnh

LightGBM	tuning_1	tuning_2	tuning_optuna
subsample	0,5	0,9	0,894
learning_rate	0,05	0,2	0,085
num_iterations	300	200	4778
max_depth	10	5	45

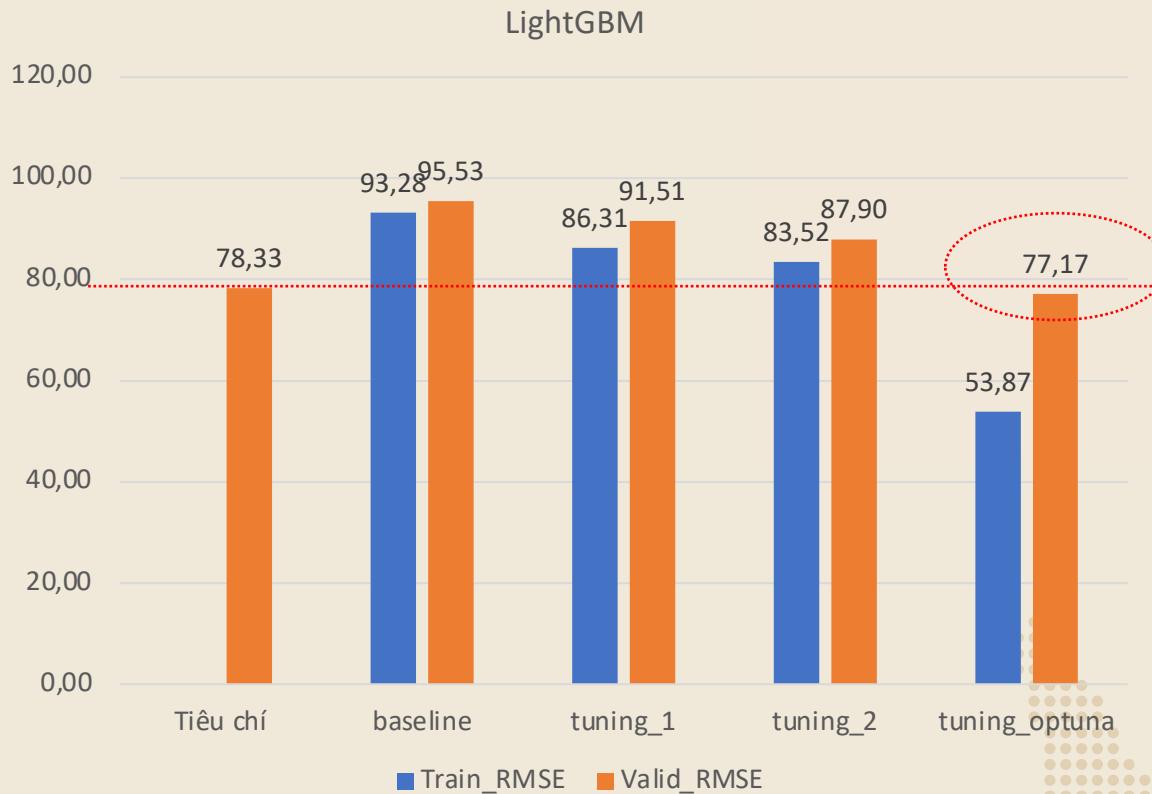
XGBoost	tuning_1	tuning_2	tuning_optuna
subsample	0,8	0,95	0,984
learning_rate	0,05	0,05	0,053
n_estimators	3000	3500	3255
max_depth	5	5	6

CatBoost	tuning_1	tuning_2	tuning_optuna
subsample	0,5	0,8	0,658
learning_rate	0,1	0,05	0,099
iterations	1000	3000	2808
depth	7	7	6

So sánh các kết quả

Kết quả của LightGBM

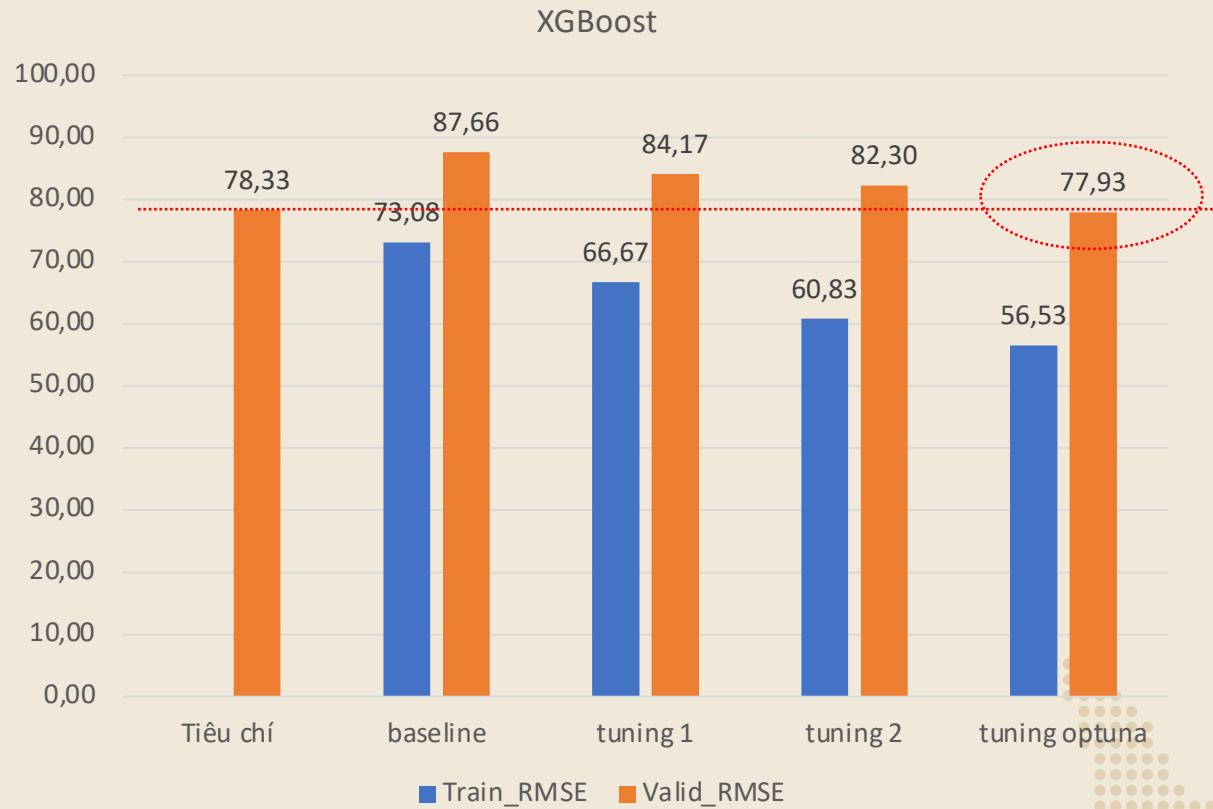
- Mô hình **đạt** tiêu chí RMSE \leq **78,83**
- Kết quả trên tập kiểm định tốt nhất khi tuning tham số bằng Optuna: **RMSE = 77,17**



So sánh các kết quả

Kết quả của XGBoost

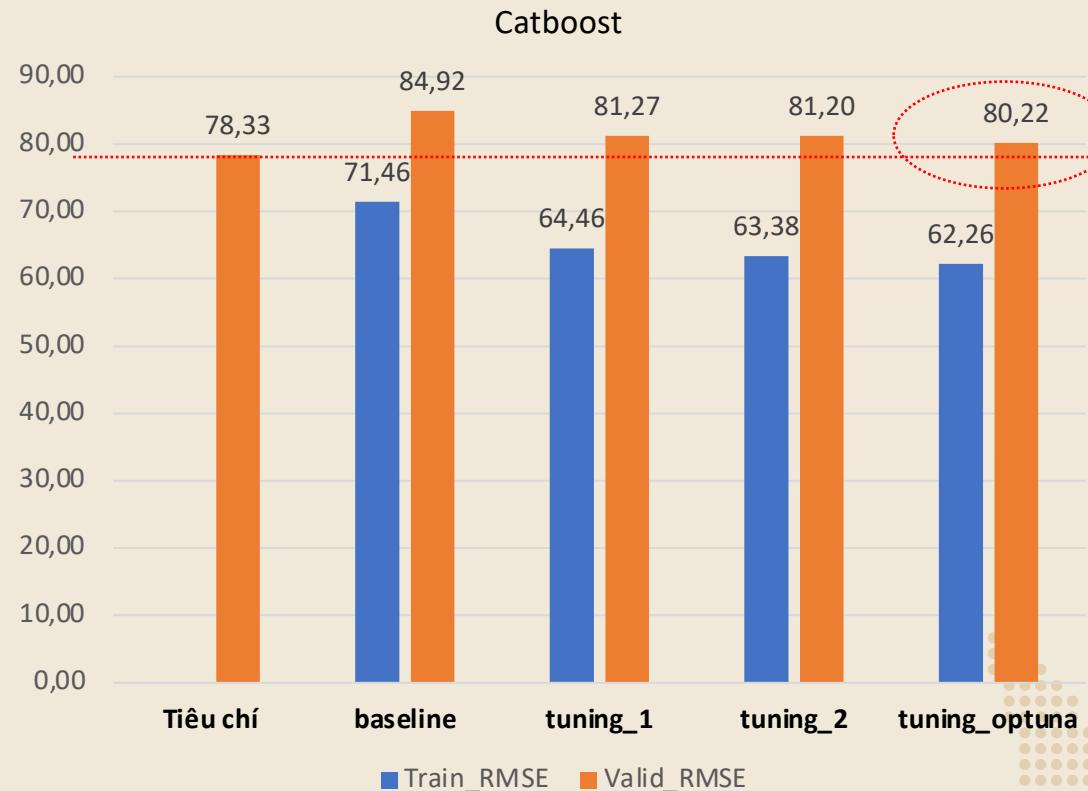
- Mô hình **đạt** tiêu chí RMSE \leq **78,83**
- Kết quả trên tập kiểm định tốt nhất khi tuning tham số bằng Optuna: **RMSE = 77,93**



So sánh các kết quả

Kết quả của CatBoost

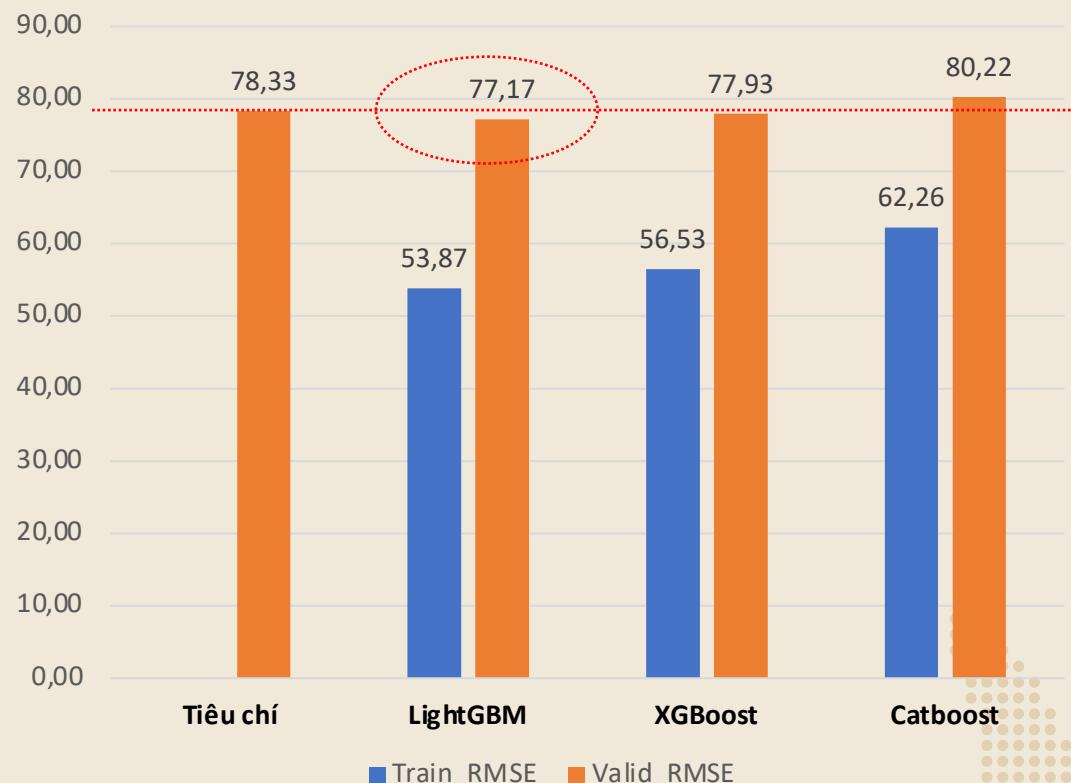
- Mô hình **chưa đạt tiêu chí**
RMSE <= 78.83
- Tuy nhiên kết quả có cải thiện khi tuning tham số
- Kết quả trên tập kiểm định tốt nhất khi tuning tham số bằng Optuna: **RMSE = 80.22**



So sánh các kết quả

Tổng hợp các kết quả

- So sánh kết quả của các mô hình tốt nhất sau khi tuning
- Đạt tiêu chí **RMSE ≤ 78.33**
- Kết quả trên tập kiểm định tốt nhất: **LightGBM (RMSE = 77.17)**



IV

Kết luận



Kết luận

Đánh giá: đạt tiêu chí đánh giá (RMSE <= 78.33) với kết quả tốt nhất là **RMSE = 77.17**

Các đặc trưng được chọn:

- 12 đặc trưng ban đầu: tourist_area, spot_facility, area, city, type, category, tourism_index, info, event, year, month, day
- 15 đặc trưng tạo mới: day_of_week, day_of_year, Weekday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday, Spring, Summer, Fall, Winter, Holiday

Mô hình cuối cùng được chọn:

- Mô hình học máy Gradient Boosting sử dụng framework **LightGBM**

Các giá trị tham số tinh chỉnh (các tham số còn lại để giá trị mặc định):

- subsample: 0,894
- learning_rate: 0,085
- num_iterations: 4778
- max_depth: 45

Source code: https://github.com/supham95/DSP305x_Final-Project_Option-1_supcfx16803



Mở rộng và định hướng trong tương lai

Nếu có cơ hội, trong tương lai em định hướng sẽ cải thiện mô hình dự đoán khách du lịch tốt hơn bằng cách:

- Sử dụng các mô hình học máy tiên tiến hơn
- Kết hợp thêm các nguồn dữ liệu bên ngoài (lịch sử tìm kiếm trên mạng, dữ liệu thời tiết,...)
- Phân tích dữ liệu theo thời gian thực để cập nhật dự đoán liên tục



Tài liệu tham khảo

- LightGBM Website. *Welcome to LightGBM's documentation!* <<https://lightgbm.readthedocs.io/en/stable/>>
- XGBoost Website. *XGBoost Documentation* <<https://xgboost.readthedocs.io/en/stable/>>
- CatBoost Website. *CatBoost Documentation* <<https://catboost.ai/en/docs/>>
- KENTAK0928 (2023). *Prediction of Tourist Arrivals* <<https://www.kaggle.com/competitions/prediction-of-tourist-arrivals/overview>>
- Aarshay Jain (2023). *Mastering XGBoost Parameter Tuning: A Complete Guide with Python Codes*
- <https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/#XGBoost_Parameters>
- MJ Bahmani (2023). *Understanding LightGBM Parameters* <<https://neptune.ai/blog/lightgbm-parameters-guide>>
- Mario Filho (2023). *CatBoost Hyperparameter Tuning Guide with Optuna* <<https://forecastegy.com/posts/catboost-hyperparameter-tuning-guide-with-optuna/>>
- Pham Minh Hoang (2020). Ensemble learning và các biến thể (P1) <<https://viblo.asia/p/ensemble-learning-va-cac-bien-the-p1-WAyK80AkKxX>>
- Brain John (2023). *When to Choose CatBoost Over XGBoost or LightGBM [Practical Guide]* <<https://neptune.ai/blog/when-to-choose-catboost-over-xgboost-or-lightgbm>>



Cảm ơn

