



**TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA CÔNG NGHỆ THÔNG TIN**



**TIỂU LUẬN CUỐI KỲ
HỌC PHẦN: KHOA HỌC DỮ LIỆU**

**DỰ ĐOÁN GIÁ VÀ PHÂN LOẠI PHÒNG KHÁCH SẠN CHO
DOANH NGHIỆP TẠI ĐÀ NẴNG**

Nhóm	1
Họ Và Tên Sinh Viên	Lớp Học Phần
Hồ Quốc Thiên Anh	21Nh12
Nguyễn Quang Sáng	

ĐÀ NẴNG, 05/2024

TÓM TẮT

Hiện nay, kinh doanh khách sạn đang dần hồi phục trở lại sau đại dịch COVID-19. Cùng với sự tăng lên của nhu cầu sử dụng, các chủ đầu tư cũng cần đặt ra các mức giá chính xác cho khách sạn của mình nhằm thu hút du khách. Vì vậy trong bài tập này, nhóm thu thập dữ liệu gồm các thuộc tính và giá của khách sạn từ các trang web du lịch. Sau đó lựa chọn các thuộc tính quan trọng, ảnh hưởng nhiều đến giá cả của khách sạn để đưa ra dự đoán về giá cả và phân loại hạng phòng dựa trên các mô hình hồi quy và phân loại. Dữ liệu được thu thập từ hơn 1000 khách sạn khác nhau trên thành phố Đà Nẵng và các mô hình triển khai đều đạt hiệu suất trên 80%, có thể làm một nguồn tham khảo đáng tin cậy cho các chủ đầu tư.

BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Tự đánh giá
Hồ Quốc Thiên Anh	<ul style="list-style-type: none">• Chuẩn hóa đặc trưng• Cài đặt mô hình• Huấn luyện và đánh giá hiệu quả mô hình	Đã hoàn thành
Nguyễn Quang Sáng	<ul style="list-style-type: none">• Thu thập và làm sạch dữ liệu• Trực quan hóa dữ liệu• Trích xuất, lựa chọn đặc trưng	Đã hoàn thành

MỤC LỤC

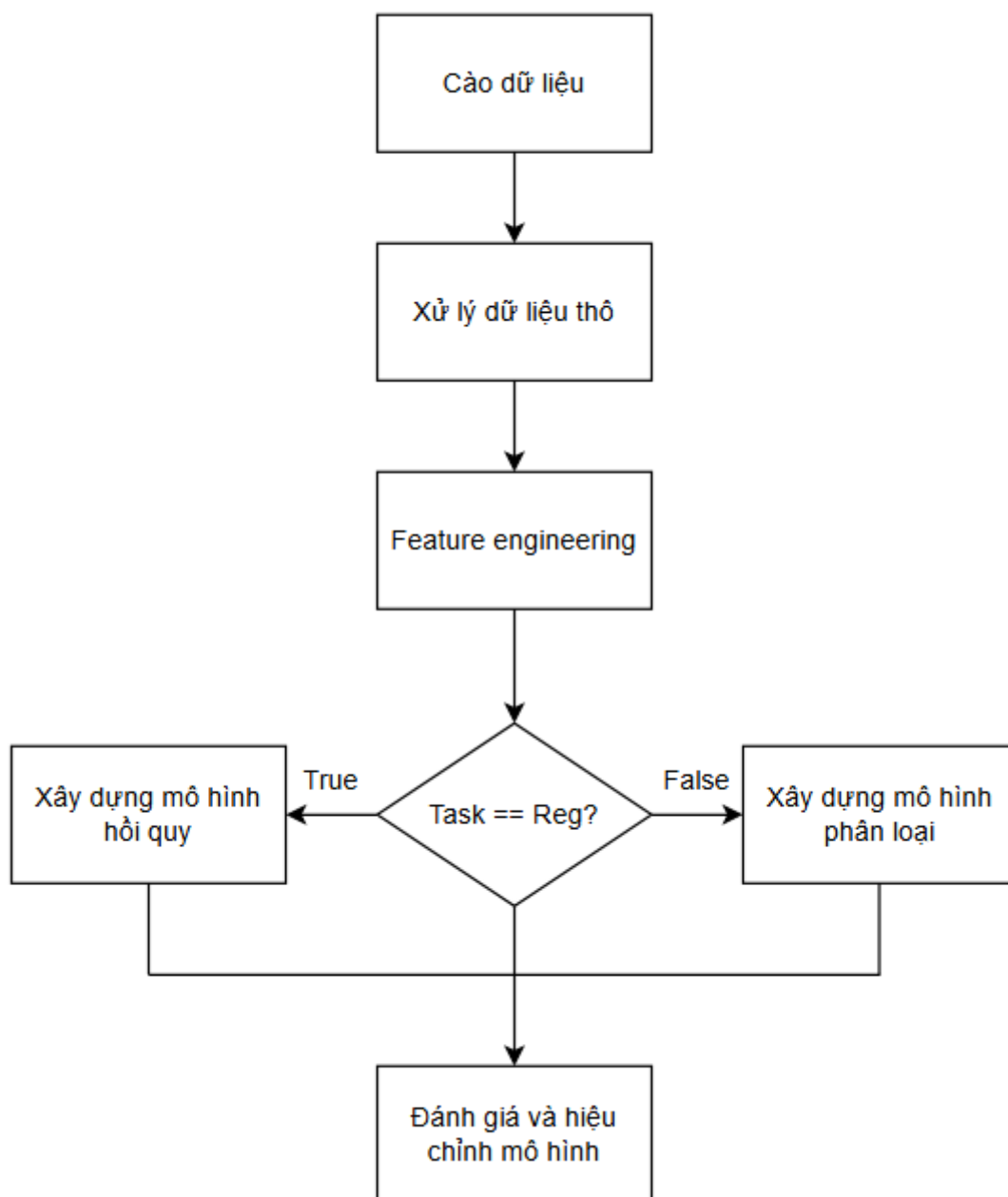
1. Giới thiệu	5
2. Thu thập và mô tả dữ liệu	6
2.1. Thu thập dữ liệu	6
2.2. Mô tả và trực quan hoá dữ liệu	8
3. Feature Engineering.....	12
3.1. Feature Extraction (Trích xuất đặc trưng).....	12
3.1.1. Sử dụng PCA	13
3.1.2. Tạo các đặc trưng mới	14
3.2. Feature Selection (Lựa chọn đặc trưng).....	14
3.2.1. Regression.....	14
3.2.2. Classification	15
3.3. Feature transformation (Biến đổi đặc trưng).....	16
3.3.1. Biến đổi phân phối của tập dữ liệu	16
3.3.2. Robust Scaling	17
4. Mô hình hóa dữ liệu.....	18
4.1. Regression	18
4.1.1. Linear Regression – Hồi quy tuyến tính	18
4.1.2. Random Forest for Regression	18
4.2. Classification.....	19
4.2.1. Softmax Regression	19
4.2.2. Random Forest for Classification	20
4.3. Kết quả thực thi mô hình.....	20
4.3.1. Kết quả bài toán Regression	20
4.3.2. Kết quả bài toán Classification	21
5. Kết luận	22
5.1. Kết quả thu được	22
5.2. Hướng phát triển	23
6. Tài liệu tham khảo	23

1. Giới thiệu

Ngày nay, với sự phát triển của ngành du lịch tại thành phố, số lượng du khách ghé thăm càng ngày càng tăng, dẫn đến nhu cầu sử dụng khách sạn ngày càng lớn. Cùng với đó, số lượng khách sạn trên địa bàn thành phố ngày một tăng nhằm đáp ứng thị trường đang dần lớn mạnh. Tuy nhiên, việc đặt các mức giá phù hợp với chất lượng của khách sạn và tệp khách hàng hướng đến không phải là điều dễ dàng đối với các chủ đầu tư.

Để giải quyết vấn đề này, nhóm đã ứng dụng các kiến thức đã học trong học phần Khoa học dữ liệu để xây dựng các mô hình dự đoán giá cả và phân loại hạng phòng với dữ liệu được thu thập trên trang web du lịch nổi tiếng Booking.com.

Giải pháp của nhóm được thể hiện trong sơ đồ khối dưới đây:



Hình 1: Pipeline chương trình

2. Thu thập và mô tả dữ liệu

2.1. Thu thập dữ liệu

2.1.1. Nguồn dữ liệu và công cụ thu thập:

- Nguồn dữ liệu được thu thập tại website: **Booking**, một trang chuyên cung cấp thông tin về nơi lưu trú (Nhà nghỉ, Khách sạn, Resort, Villa,...), đặt phòng và vé máy bay trên khắp thế giới. Đường link dẫn tới website: <https://www.booking.com>
- Dựa trên quá trình phân tích về cấu trúc và đặc trưng của website, nhóm lựa chọn các công cụ sau để tiến hành cào dữ liệu: ngôn ngữ Python và các thư viện hỗ trợ bao gồm Selenium và BeautifulSoup.

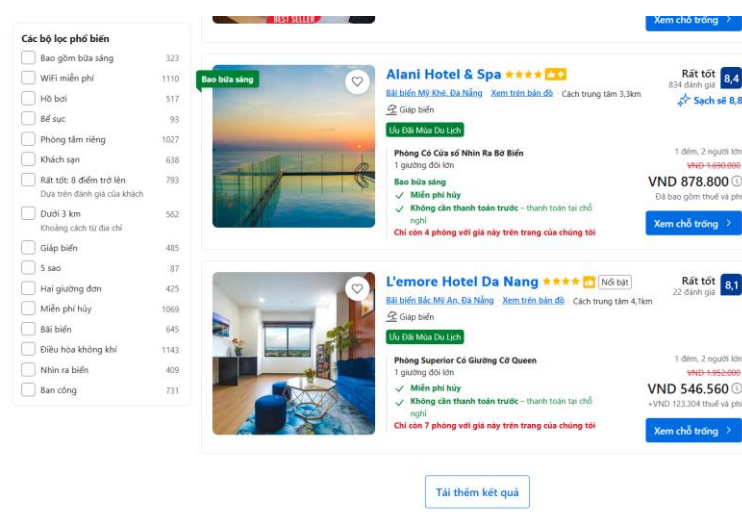
Công cụ thu thập:

- **Selenium**: Selenium là một công cụ tự động hóa, giúp giả lập các thao tác của người dùng trên trình duyệt. Nhóm sử dụng Selenium để tiến hành mở website, đợi website load dữ liệu, cuộn trang, nhấn nút load thêm dữ liệu.
- **BeautifulSoup**: Thư viện của Python, giúp phân tách source text HTML lấy được công cụ Selenium và truy xuất dữ liệu từ các tệp đó.

- Tổng quan về quá trình cào dữ liệu từ trang **Booking**:

Bước 1: Load toàn bộ dữ liệu tìm kiếm của trang **Booking**:

- Tiến hành cuộn đến cuối trang, dữ liệu khách sạn sẽ liên tục được load cho tới khi xuất hiện nút “**Tải thêm kết quả**”
- Bấm nút và cuộn trang, quá trình này lặp lại cho tới khi dữ liệu được load hoàn toàn.

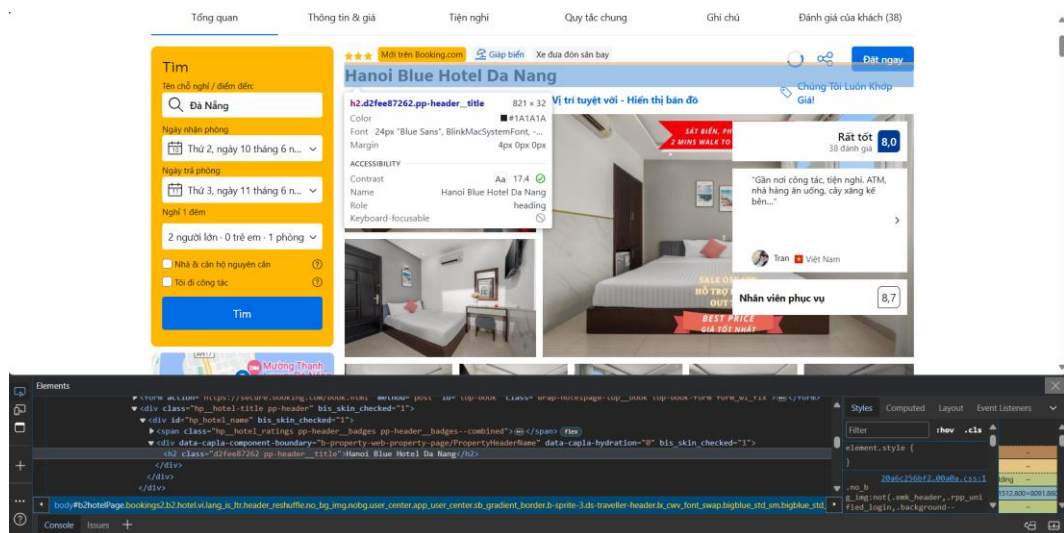


Hình 2: Trang thông tin các khách sạn

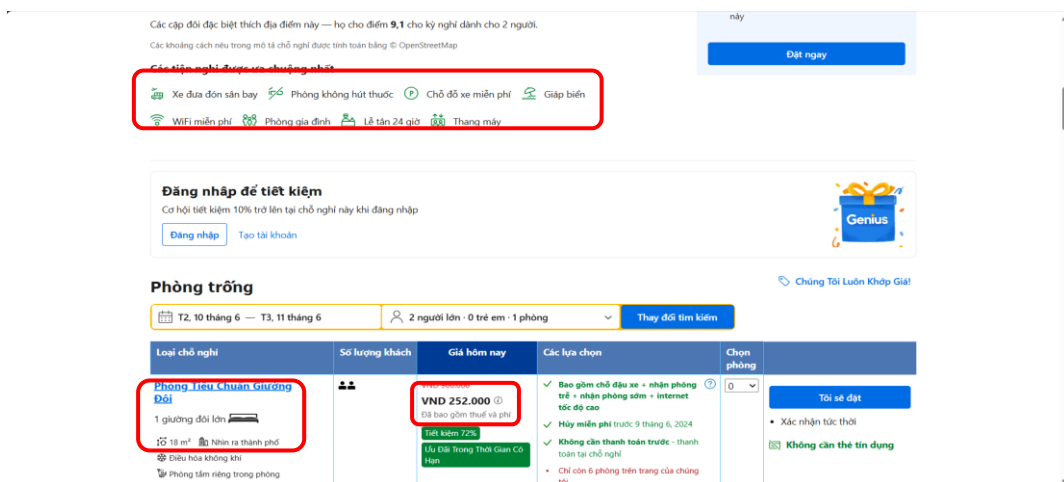
- Theo quan sát của nhóm, trang hiển thị tầm 50 khách sạn cho trang đầu tiên và 25 khách sạn tiếp theo cho lần nhấn nút “**Tải thêm kết quả**”, tổng số Khách sạn nhóm cần thu thập là > 1000 mẫu, ước tính số lần cần nhấn nút là: ≈ 40 lần nhấn. Điểm yếu của phương pháp này là dễ gây ra hiện tượng tràn RAM, đơ máy ảnh hưởng đến quá trình crawl data.

Bước 2: Tiến hành truy cập vào trang trang thông tin chi tiết của từng Khách sạn:

- Sử dụng Selenium, truy cập các trang thông tin chi tiết của từng Khách sạn, trên một Tab mới.
- Trong các trang này, bằng cách kiểm tra Class, Id, ... của thẻ thông qua tính năng **Inspect** trên trang web kết hợp với việc sử dụng công cụ **BeautifulSoup** ta có thể thu thập được các thông tin về **Tên, Địa chỉ, Giá** của Khách sạn và các thông tin liên quan có ảnh hưởng đến Giá như: **Tiện nghi, Size phòng (Phòng tiêu chuẩn), Khoảng cách đến các bãi biển, Sân bay.**



Hình 3 Thông tin vùng dữ liệu cần thu thập



Hình 4 Dữ liệu cần thu thập

Xung quanh khách sạn

Vị trí tuyệt vời - Hiển thị bản đồ

Xem phòng trống

Xung quanh có gì?

Cong Vien Bien Pham Van Dong	950 m
pullman Playground	1,7 km
Cầu khóa Tình yêu Đà Nẵng	1,8 km
Cầu Trần Thị Lý	1,8 km
Riverside Park	1,8 km
Upside Down World Danang	1,8 km
Cầu Rồng	2,1 km
Apec Park	2,3 km
Bảo tàng điêu khắc Chăm	2,4 km
Garuda Valley	2,5 km

Nhà hàng & quán cà phê

Cafe/quán bar · Ca Phe Bien Nho	30 m
Cafe/quán bar · Ca Phe Thanh Thanh	30 m
Nhà hàng · Mi Quảng Bà Dung	30 m

Địa điểm tham quan hàng đầu

Cầu sông Hàn	2,6 km
Thuan Phuoc Field	5 km

Cảnh đẹp thiên nhiên

Núi · Ngũ Hành Sơn/ Núi Non Nước	2,1 km
Đỉnh núi · Núi Sơn Trà	8 km

Các bãi biển trong khu vực

Bãi biển Mỹ Khê	300 m
Bãi biển Bắc Mỹ An	1,2 km
Bãi biển Non Nước	4,9 km
Bãi biển Thanh Bình	5 km
Bãi Bụt	5 km

Phương tiện công cộng

Xe buýt · Bến xe Buýt	7 km
-----------------------	------

Các sân bay gần nhất

Sân bay Quốc tế Đà Nẵng	4,3 km
Sân bay quốc tế Phú Bài	67 km
Sân bay Chu Lai	83 km

Tất cả khoảng cách được đo theo đường thẳng. Khoảng cách di chuyển thực sự có thể khác.



Bạn không tìm thấy một số thông tin? [Đúng vậy](#) / [Không phải](#)

Hình 5 Thông tin về khoảng cách

- Thông qua các đánh giá khảo sát từ các khách sạn, **Feature Tiện Nghi** sẽ được chọn lọc để thu thập, bao gồm: Hồ bơi, Chỗ đỗ xe, Phòng không hút thuốc, Giáp biển, WiFi miễn phí, Phòng gia đình, Quầy Bar và Buổi ăn sáng.
- Khoảng cách đến Sân bay và Khoảng cách đến Biển được thu thập với giá trị nhỏ nhất trong các khoảng cách cùng loại.

Bước 3: Đóng Tab chi tiết Khách sạn và tiếp tục mở Tab chi tiết của khách sạn tiếp theo

- Sau khi thu thập toàn bộ thông tin cần thiết, sử dụng công cụ Selenium để đóng Tab thông tin khách sạn hiện tại và tiếp tục tạo Tab mới cho Khách sạn tiếp theo để tiến hành thu thập
- Quá trình này lặp đi lặp lại cho đến khách sạn cuối cùng

Name	Date modified	Type	Size
 raw_data_test.csv	5/21/2024 4:02 PM	Microsoft Excel C...	14 KB
 raw_data_train.csv	4/8/2024 11:38 PM	Microsoft Excel C...	132 KB

Hình 6: Thời gian crawl dữ liệu

2.2. Mô tả và trực quan hoá dữ liệu

Sau khi tiến hành crawl data, nhóm thu được tập dữ liệu với:

- Số lượng mẫu cho tập Train và Valid: 1011 mẫu.
- Số lượng mẫu cho tập Test: 102 mẫu
- Số lượng đặc trưng ở 1 mẫu: 14

Sau khi có dữ liệu thô, tiến hành làm sạch dữ liệu bằng cách:

- Xóa đi các kí tự thừa: Xóa chuỗi kí tự VNĐ ở phần đuôi của Feature Price
- Ép kiểu dữ liệu sang kiểu dữ liệu thích hợp

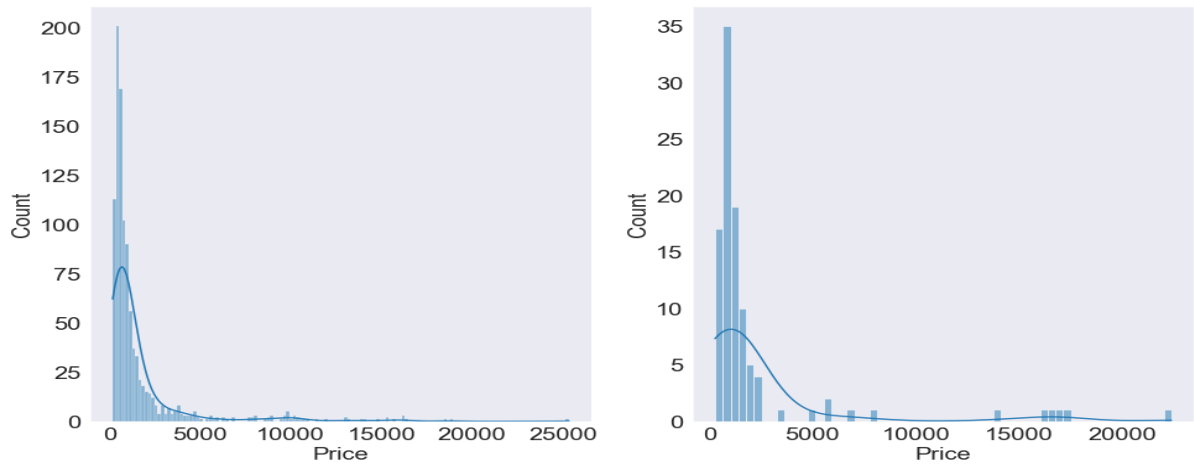
Kết quả thu được tập dữ liệu sau làm sạch được mô tả qua 2 bảng dưới đây.

Bảng 1: Mô tả tổng quan 2 tập dữ liệu

	Số lượng đặc trưng	Số lượng mẫu
Tập train	14	1011
Tập test	14	102

Bảng 2: Mô tả cụ thể tập dữ liệu

STT	Đặc trưng	Mô tả	Kiểu dữ liệu	Số mẫu dữ liệu trống
1	Name	Tên khách sạn	String	0
2	Address	Địa chỉ	String	0
3	Price	Giá phòng (Nghìn VND)	Float	0
4	Size	Kích thước phòng (m ²)	Int	0
5	Distance to beach	Khoảng cách tới biển (km)	Float	0
6	Distance to airport	Khoảng cách tới sân bay (km)	Float	0
7	Pool	Có/Không có hồ bơi	Boolean	0
8	Bar	Có/Không quầy bar	Boolean	0
9	Car	Có/Không bãi đỗ xe miễn phí	Boolean	0
10	Non-smoking room	Có/Không phòng không hút thuốc	Boolean	0
11	Near beach	Gần biển	Boolean	0
12	WiFi	Có/Không wifi miễn phí	Boolean	0
13	Family room	Có/Không phòng gia đình	Boolean	0
14	Breakfast	Có/Không bữa sáng miễn phí	Boolean	0

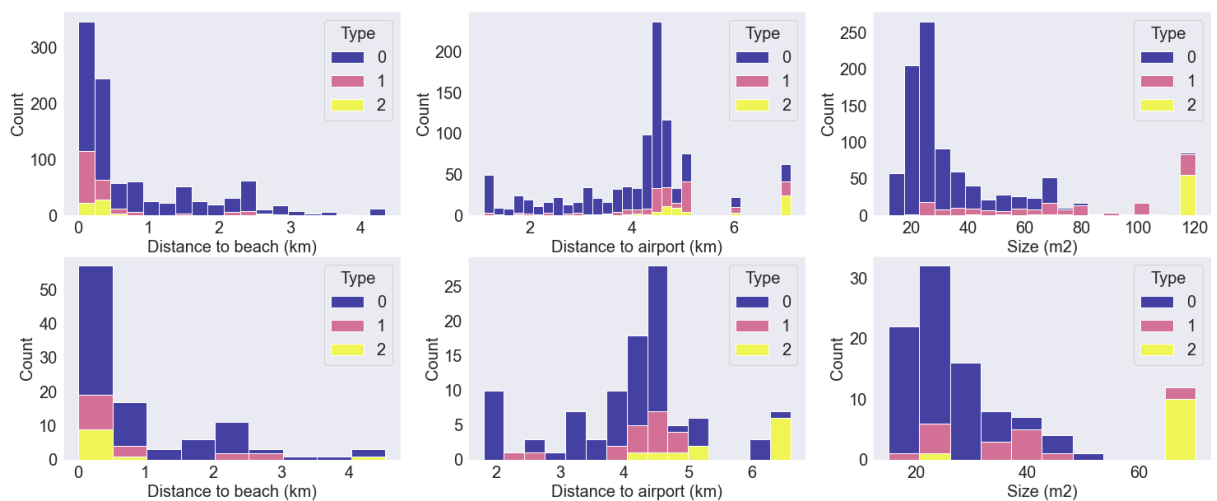


Hình 7: Phân bố dữ liệu của biến Price trên tập train (trái) và test (phải)

- Giá phòng khách sạn trên thị trường hiện nay chủ yếu rơi vào khoảng 250.000 – 1.500.000 VND. Tuy nhiên, có thể thấy rằng trong thành phố Đà Nẵng vẫn có những khách sạn cao cấp với mức giá từng đêm lên đến vài chục triệu đồng. Từ nhận xét này, ta có thể chia các hạng phòng khách sạn ở Đà Nẵng thành 3 mức như sau:

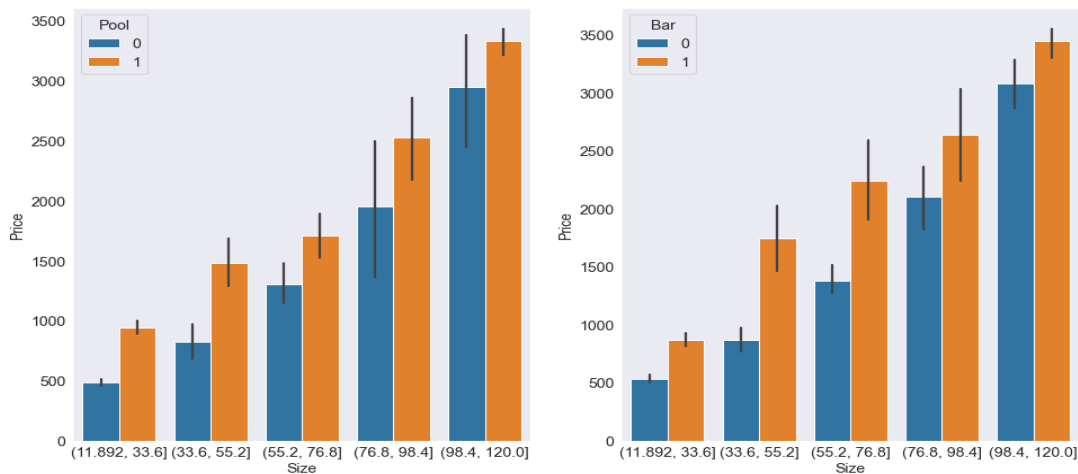
- Dưới 1.500.000 VND: hạng thường
- 1.500.000 – 5.000.000 VND: hạng cao cấp
- Trên 5.000.000 VND: hạng sang trọng

- Ta có thể chia như trên bởi phần lớn khách sạn có giá phòng rẻ hơn 1.500.000 VND và từ biểu đồ ta có thể thấy sự phân phối của giá phòng có sự thay đổi lớn từ mốc 5.000.000 VND. Ngoài ra phân chia bằng 2 mức giá này còn giúp đảm bảo số lượng mẫu ở mỗi mức vừa đủ cho việc huấn luyện mô hình.



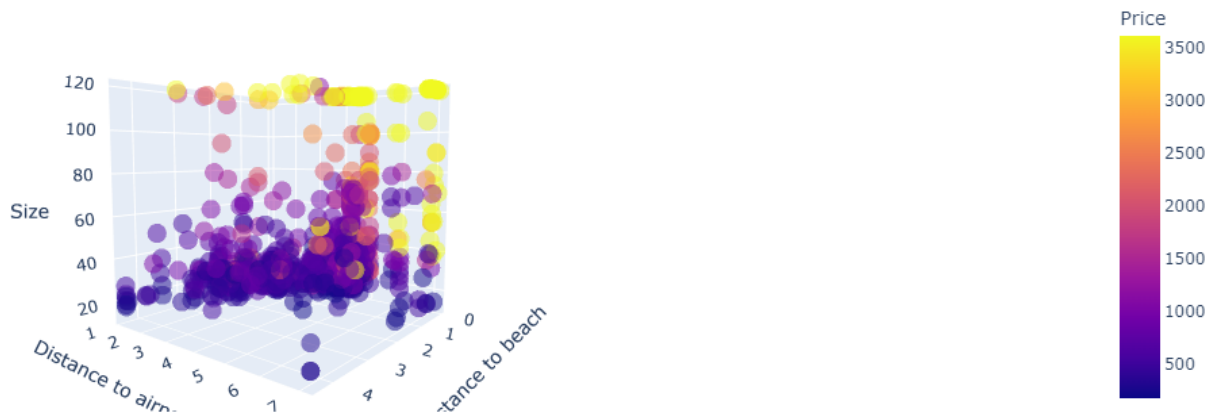
Hình 8: Phân bố dữ liệu của các biến số thực trên tập train (trên) và test (dưới)

- Từ phân bố dữ liệu của 2 biến distance, ta có thể thấy được phần lớn khách sạn ở thành phố Đà Nẵng được xây về hướng gần biển hơn so với xây gần sân bay. Và hầu hết các khách sạn giá cao đều được xây gần biển.
- Tập dữ liệu của cả tập train lẫn test đều có rất nhiều ngoại lệ trong cả 4 biến số thực. Nguyên nhân của việc này chính là vì sự tồn tại của các khách sạn cao cấp và đặc biệt cao cấp ở trên địa bàn thành phố với các mức giá phòng đắt đỏ.
- Ngoài ra, ta có thể thấy rằng không có sự thay đổi quá lớn trong phân phối dữ liệu của các biến ở trên 2 tập huấn luyện và tập kiểm thử. Điều này sẽ giúp ích cho việc đảm bảo hiệu suất mô hình sau này. Riêng đối với lĩnh vực khách sạn, chúng ta sẽ không quá lo lắng việc mô hình sẽ gặp phải các tập dữ liệu test có phân phối dữ liệu bất thường (trong điều kiện cùng vị trí địa lý)



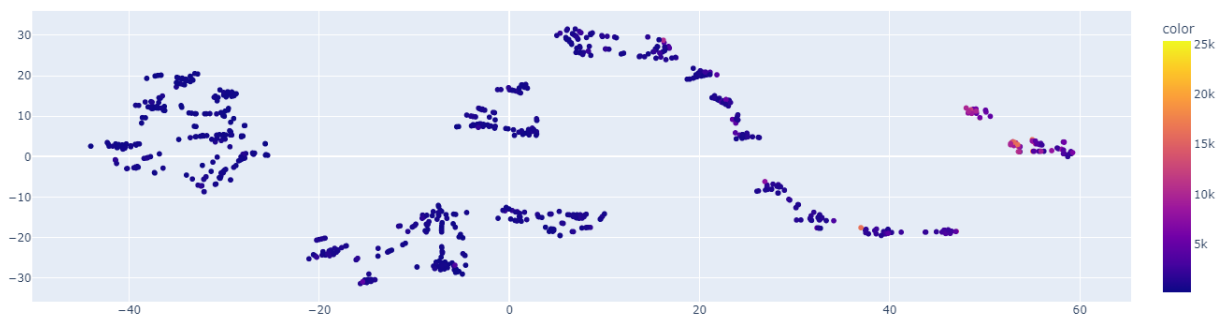
Hình 9: Biểu đồ khảo sát ảnh hưởng của Pool và Bar đến giá phòng

- Từ biểu đồ trên, ta có thể thấy trong cùng 1 phân khúc về kích thước phòng, các khách sạn có cung cấp dịch vụ hồ bơi và quầy bar sẽ có giá thuê cao hơn. Các chủ khách sạn nếu muốn tăng giá phòng có thể cân nhắc việc cung cấp 2 loại dịch vụ này.



Hình 10: Biểu đồ tương quan giữa các biến distance và size đến giá phòng

- Từ biểu đồ trên, ta có thể rút ra được các kết luận như sau:
 - Phòng càng lớn giá càng cao.
 - Các khách sạn gần biển thường sẽ có giá phòng cao hơn so với các khách sạn gần sân bay.
 - Kích thước phòng có ảnh hưởng đối với giá phòng nhiều hơn so với 2 biến còn lại.



Hình 11: Sử dụng t-SNE để tìm quy luật cụm của dữ liệu

- Từ biểu đồ, ta thấy tồn tại 2 cụm dữ liệu hoàn toàn đối lập nhau.
- Vùng dữ liệu ở giữa không thể phân cụm rõ, điều này sẽ gây khó khăn cho mô hình trong việc phân loại các khách sạn ở các vị trí này.

3. Feature Engineering

3.1. Feature Extraction (Trích xuất đặc trưng)

- Nhóm áp dụng 2 kỹ thuật trích xuất đặc trưng sau cho cả 2 bài toán Regression và Classification:

- Sử dụng PCA để giảm chiều dữ liệu
- Tạo các đặc trưng mới từ các đặc trưng cũ

3.1.1. Sử dụng PCA

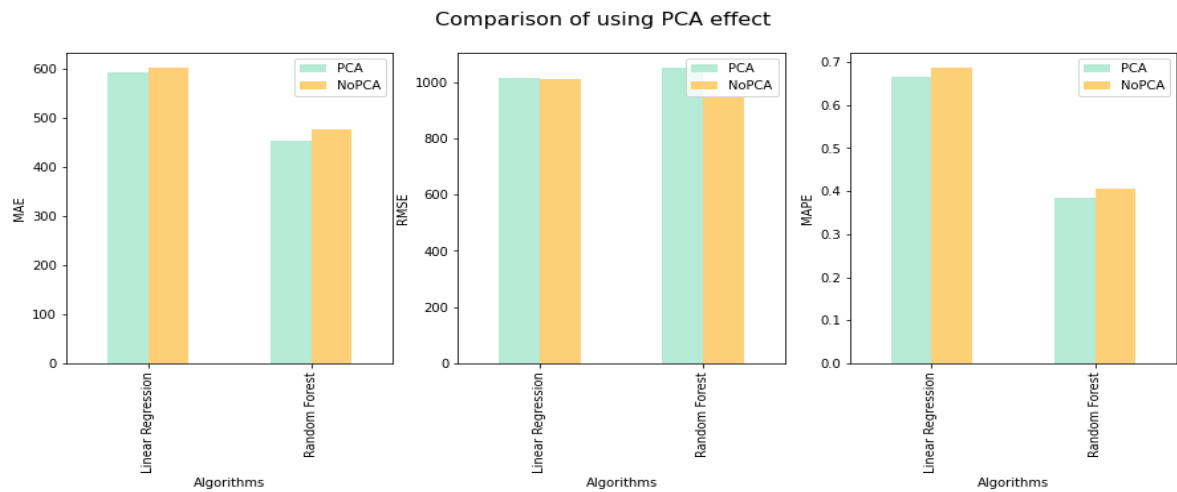
- Trong bài toán Regression, 2 thuật toán nhóm sẽ sử dụng là Linear Regression và Random Forest. Đối với bài toán Classification, 2 thuật toán sử dụng sẽ là Softmax Regression và Random Forest

- Để đánh giá sơ bộ hiệu quả của các công việc trên, nhóm sử dụng các metrics đối với từng bài toán như sau:

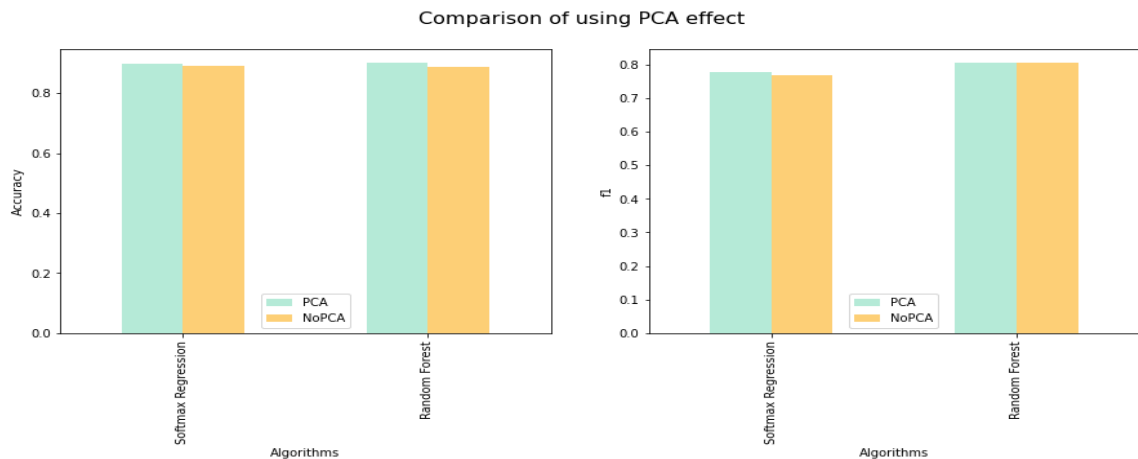
- Regression: MAE, RMSE và MAPE

- Classification: Accuracy và F1 score

- Chạy thử nghiệm mô hình khi áp dụng PCA và không áp dụng PCA trên tập huấn luyện, ta thu được biểu đồ như dưới đây:



Hình 12: Biểu đồ khảo sát hiệu suất mô hình trước và sau khi sử dụng PCA với bài toán hồi quy



Hình 13: Biểu đồ khảo sát hiệu suất mô hình trước và sau khi sử dụng PCA với bài toán phân loại

- Từ biểu đồ ta có thể thấy việc sử dụng PCA không cải thiện đáng kể hiệu suất mô hình, vậy nên ta sẽ không sử dụng PCA cho cả 2 bài toán.

3.1.2. Tạo các đặc trưng mới

- Trong bài toán này, ta sẽ tạo thêm 2 biến mới:

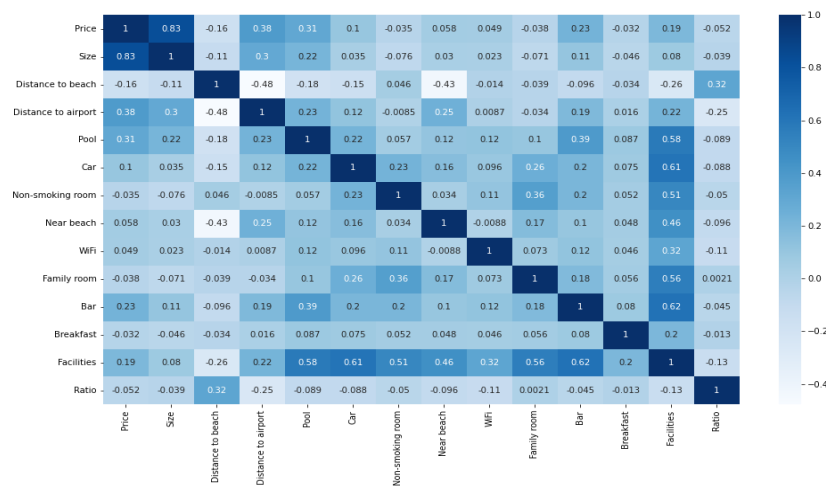
- Ratio: tỉ lệ giữa 2 biến Distance to beach và Distance to airport.
- Facilities: tổng số lượng các dịch vụ được ưa thích mà khách sạn cung cấp

- Chúng ta sẽ đánh giá mức độ quan trọng của 2 biến này trong bước lựa chọn đặc trưng.

3.2. Feature Selection (Lựa chọn đặc trưng)

3.2.1. Regression

- Để lựa chọn đặc trưng cho thuật toán Linear Regression, ta sẽ dựa vào biểu đồ tương quan giữa các đặc trưng

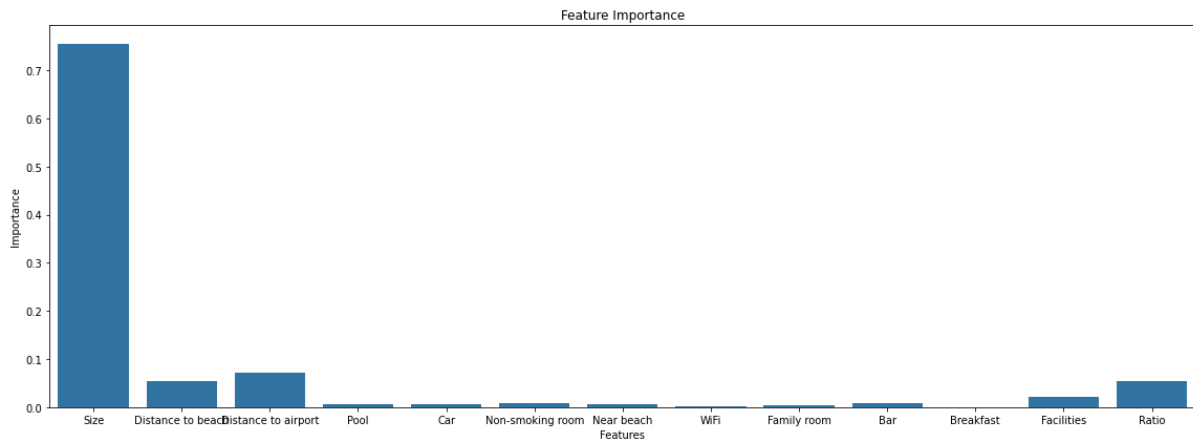


Hình 14: Biểu đồ tương quan giữa các đặc trưng

- Từ biểu đồ ta có thể thấy được top 4 đặc trưng quan trọng nhất cho Linear Regression là:

- Size
- Distance to airport
- Distance to beach
- Facilities

- Đối với thuật toán Random Forest, ta sẽ sử dụng thuộc tính feature_importance để chọn lọc



Hình 15: Biểu đồ mức độ quan trọng của các đặc trưng trong thuật toán Random Forest cho hồi quy

- Từ biểu đồ ta có top 4 đặc trưng quan trọng cho thuật toán Random Forest là:

- Size
- Distance to airport
- Distance to beach
- Ratio

- Ngoài ra từ kết quả EDA, ta sẽ chọn thêm 2 biến Pool và Bar cho 2 mô hình hồi quy

=> 2 mô hình hồi quy sẽ có tổng cộng 7 đặc trưng: Size, Distance to airport, Distance to beach, Facilities, Ratio, Pool, Bar.

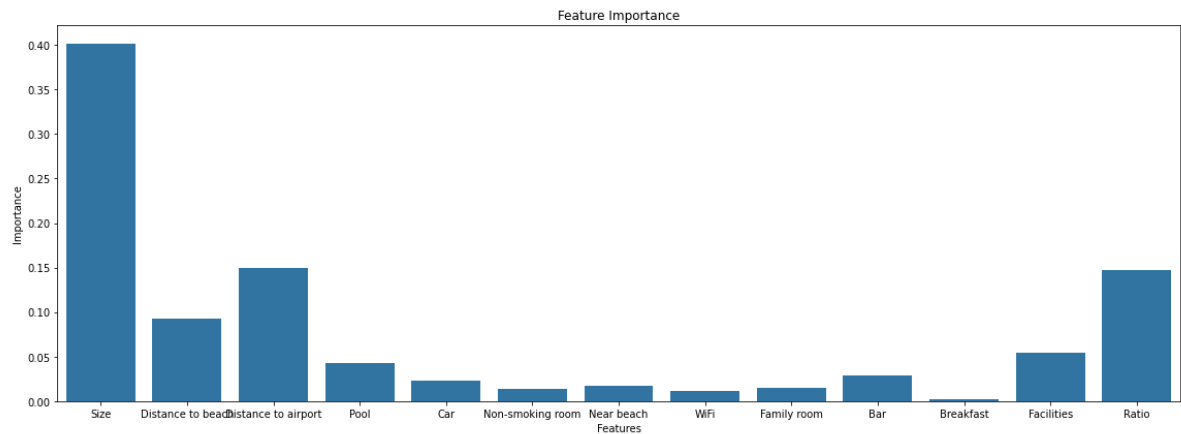
3.2.2. Classification

- Đối với thuật toán Softmax Regression, ta sử dụng Lasso Regression để loại bỏ các đặc trưng không quan trọng

- Từ đó ta chọn được 3 đặc trưng là :

- Size
- Distance to airport
- Ratio

- Đối với thuật toán Random Forest, ta cũng áp dụng kỹ thuật tương tự như ở bài toán Regression với biểu đồ mức độ quan trọng như sau :



Hình 16: Biểu đồ mức độ quan trọng của các đặc trưng trong thuật toán Random Forest cho phân loại

- Ta thấy top 5 đặc trưng quan trọng nhất cho mô hình là:

- Size
- Distance to beach
- Distance to airport
- Ratio
- Facilities

- Ta cũng sử dụng thêm 2 đặc trưng Pool và Bar dựa trên kết quả EDA

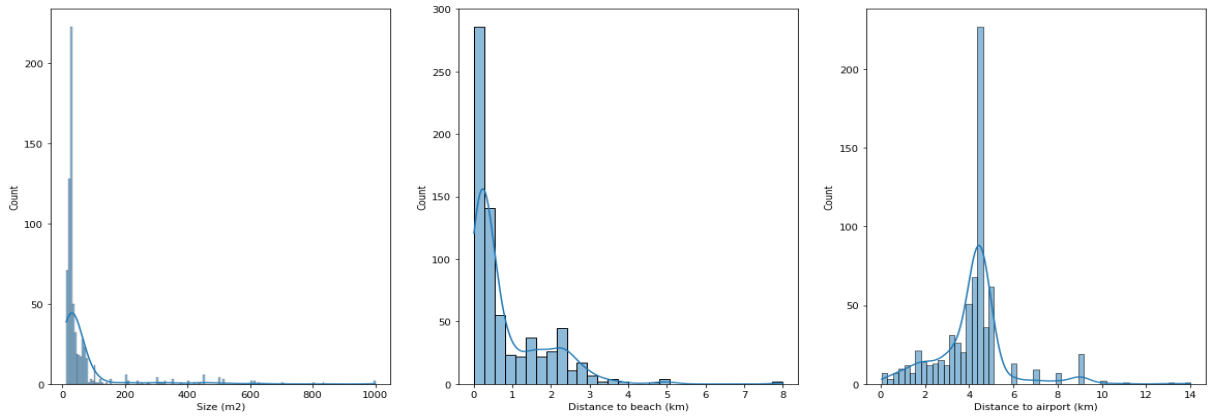
=> Như vậy, 2 mô hình phân loại cũng sẽ sử dụng chung bộ đặc trưng với 2 mô hình hồi quy.

3.3. Feature transformation (Biến đổi đặc trưng)

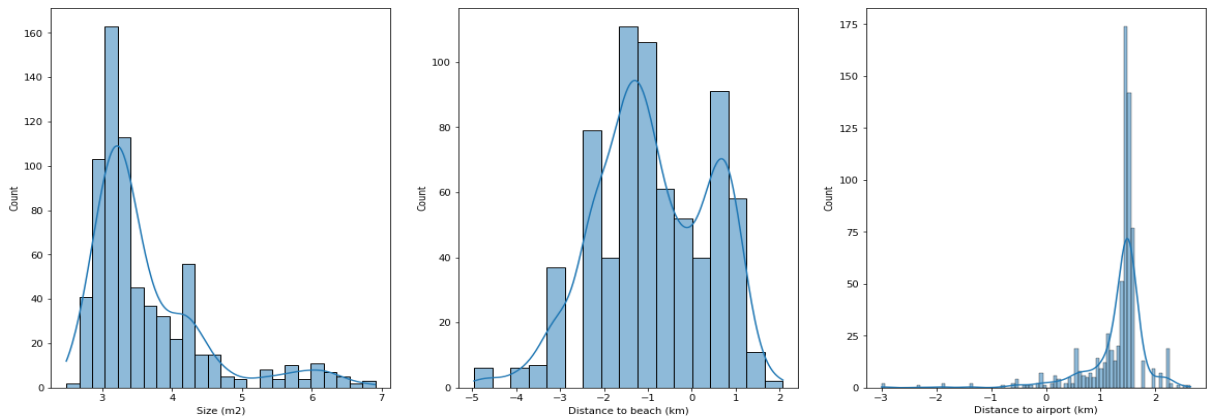
- Riêng ở bước biến đổi đặc trưng, các bước thao tác sẽ hoàn toàn giống nhau ở cả 2 bài toán Regression và Classification

3.3.1. Biến đổi phân phối của tập dữ liệu

- Ở mục 2, ta thấy phân phối dữ liệu của cả 4 biến số thực đều bị lệch trái, điều này sẽ phần nào ảnh hưởng đến hiệu suất của mô hình. Vì vậy, chúng ta cần biến đổi phân phối dữ liệu của các biến này thành phân phối chuẩn bằng cách lấy logarit của từng giá trị.



Hình 17: Phân phối dữ liệu của 3 biến dự đoán trước khi biến đổi



Hình 18: Phân phối dữ liệu của 3 biến dự đoán sau khi biến đổi

3.3.2. Robust Scaling

- Vì tập dữ liệu chứa rất nhiều ngoại lệ, nên Robust Scaling sẽ thích hợp nhất trong việc chuẩn hóa khoảng giá trị của tập dữ liệu.
- Mặc dù sau khi thực hiện biến đổi phân phối dữ liệu, khoảng giá trị của các biến đã cải thiện đáng kể nhưng thực nghiệm cho thấy áp dụng Robust Scaling có cải thiện hiệu suất mô hình

Bảng 3: Bảng đánh giá hiệu suất mô hình trước khi Robust Scaling

Thuật toán	MAE	RMSE	MAPE
Linear Regression	910.50	1272.03	1.09
Random Forest	456.11	901.79	0.39

- Sau khi thực hiện Robust Scaling

Bảng 4: Bảng đánh giá hiệu suất mô hình sau khi Robust Scaling

Thuật toán	MAE	RMSE	MAPE
Linear Regression	905.17	1262.12	1.11
Random Forest	436.01	833.66	0.37

- Ta cũng nhận được kết quả tương tự khi khảo sát với bài toán Classification

4. Mô hình hóa dữ liệu

- Ta sẽ chia tỉ lệ cho tập Huấn luyện/Xác thực/Kiểm thử theo tỉ lệ 70/20/10, số lượng cụ thể như sau:

Bảng 5: Bảng mô tả số mẫu dữ liệu trong mỗi tập dữ liệu

Tổng số mẫu dữ liệu	1113
Số mẫu huấn luyện	808
Số mẫu xác thực	203
Số mẫu kiểm thử	102

- Sử dụng GridSearchCV để tìm bộ siêu tham số tốt nhất cho thuật toán Random Forest cũng như thực hiện cross-validation cho mô hình.

4.1. Regression

4.1.1. Linear Regression – Hồi quy tuyến tính

- Hồi quy tuyến tính được xây dựng dựa trên giả định rằng quan hệ giữa biến độc lập và biến phụ thuộc có thể được miêu tả bằng một đường thẳng. Vì vậy, nhiệm vụ của chúng ta là tìm được một đường thẳng tốt nhất để biểu diễn mô hình này sao cho khoảng cách giữa các điểm dữ liệu thực tế và điểm dữ liệu dự đoán trên đường thẳng là nhỏ nhất.

- Phương trình tuyến tính của mô hình (vectorized form):

$$\hat{y} = \mathbf{w} \cdot \mathbf{X} + \mathbf{b}$$

Trong đó:

- \hat{y} là vector chứa giá trị dự đoán.
- \mathbf{X} là ma trận đặc trưng, với số hàng là số lượng mẫu, số cột là số lượng đặc trưng
- \mathbf{w} : vector trọng số ứng với từng phần tử trong \mathbf{X} .
- \mathbf{b} : vector hệ số tự do, có số phần tử bằng số lượng mẫu.

- Mô hình không có bộ siêu tham số để điều chỉnh.

4.1.2. Random Forest for Regression

- Random Forest là một thuật toán học máy được sử dụng cho các bài toán hồi quy (Regression) và phân loại (Classification). Được xây dựng dựa trên ý tưởng sử dụng nhiều cây quyết định (Decision tree) khác nhau để dự đoán kết quả, từ đó giúp giảm hiện tượng Overfitting và tăng tính tổng quát cho mô hình.

- **Mục tiêu:** Dùng để dự đoán giá trị liên tục

- Cách thức hoạt động:

- Xây dựng nhiều cây quyết định (Decision Tree) độc lập.
- Mỗi cây được huấn luyện trên một tập dữ liệu con (Bootstrap Sample) và một phần của các đặc trưng (Random Feature Subset).
- Dự đoán cuối cùng là trung bình của dự đoán từ tất cả các cây (trung bình các giá trị dự đoán của các cây).

- Bộ siêu tham số:

- *n_estimators*: [100, 200, 300]: số lượng cây trong thuật toán
- *max_depth*: [None, 10, 20]: độ sâu tối đa của mỗi cây
- *min_samples_split*: [1, 2, 4]: số lượng mẫu tối thiểu cần để chia một nút trong
- *min_samples_leaf*: [1, 2, 4]: số lượng mẫu tối thiểu cần để tạo nút lá
- *max_features*: [None, 1, 2]: số lượng đặc trưng tối đa được xem xét khi tạo cây
- *bootstrap*: [True, False]: có thực hiện lấy mẫu ngẫu nhiên không

- Sử dụng GridSearchCV để thu được bộ siêu tham số tối ưu:

- *n_estimator*: 200
- *max_depth*: None
- *min_samples_split*: 2
- *min_sample_leaf*: 4
- *max_features*: None
- *bootstrap*: True

4.2. Classification

4.2.1. Softmax Regression

Softmax Regression là một thuật toán học máy được sử dụng cho bài toán phân loại đa lớp (multi-class classification). Khác với Logistic Regression, Softmax Regression là phiên bản tổng quát hơn, được sử dụng để phân loại dữ liệu vào nhiều lớp khác nhau.

- Công thức hàm Softmax

$$a_i = \frac{\exp(z_i)}{\sum_{j=1}^C \exp(z_j)}, \forall i = 1, 2, \dots, C$$

Trong đó:

- a_i là xác suất để input x rơi vào Class i
- z_i là giá trị đầu ra của Class i

- C là số lượng Class cần phân loại

- Không có bộ siêu tham số để điều chỉnh

4.2.2. Random Forest for Classification

- **Mục tiêu:** Dùng để phân loại dữ liệu đầu vào của các lớp khác nhau

- **Cách hoạt động:**

- Xây dựng nhiều cây quyết định (Decision Tree) độc lập.
- Mỗi cây được huấn luyện trên một tập dữ liệu con (Bootstrap Sample) và một phần của các đặc trưng (Random Feature Subset).
- Dự đoán cuối cùng là lớp phổ biến nhất (lớp có số phiếu bầu cao nhất).

- Bộ siêu tham số: tương tự như ở mục 4.1.2

- Sau khi sử dụng GridSearchCV, ta thu được bộ siêu tham số tối ưu như sau:

- `n_estimators`: 100
- `max_depth`: 30
- `min_samples_split`: 2
- `min_samples_leaf`: 2
- `max_features`: 3
- `bootstrap`: True

4.3. Kết quả thực thi mô hình

4.3.1. Kết quả bài toán Regression

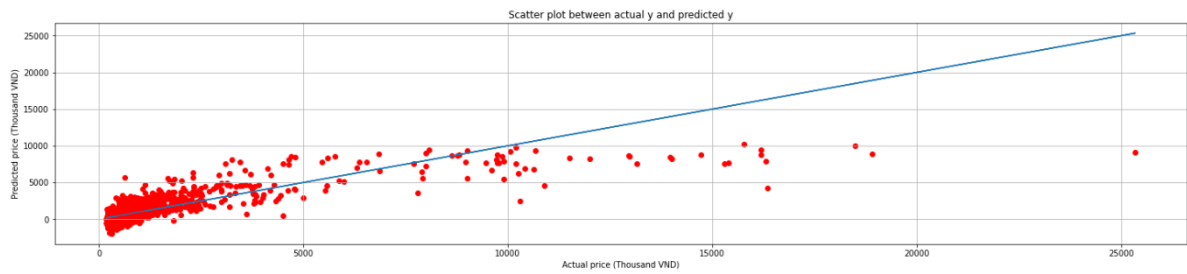
- Nhóm sử dụng 4 metrics : MAE, RMSE, MAPE và R2 score để đánh giá hiệu suất mô hình, kết hợp với đồ thị scatterplot thể hiện sự tương quan giữa giá trị dự đoán và giá trị thực tế

Bảng 6: Bảng đánh giá hiệu suất mô hình trên tập xác thực

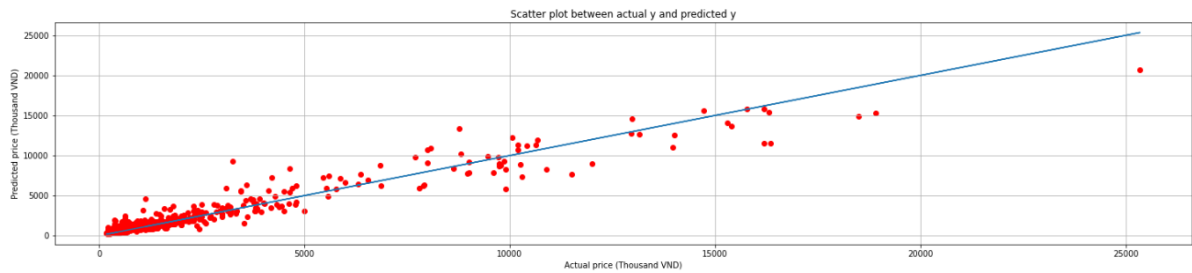
Thuật toán	MAE	RMSE	MAPE	R2 score
Linear Regression	905.17	1262.12	1.12	0.70
Random Forest	440.44	826.87	0.37	0.87
Random Forest sau khi fine-tuned	427.27	841.78	0.35	0.87

Bảng 7: Bảng đánh giá hiệu suất mô hình trên tập kiểm thử

Thuật toán	MAE	RMSE	MAPE	R2 score
Linear Regression	967.54	1560.87	1.08	0.64
Random Forest	310.99	685.60	0.24	0.93



Hình 18: Biểu đồ tương quan giữa giá dự đoán và giá thực tế (Linear Regression)



Hình 19: Biểu đồ tương quan giữa giá dự đoán và giá thực tế (Random Forest)

- Từ các kết quả trên ta thu được các kết luận sau:

- Mô hình Random Forest cho kết quả cao hơn Linear Regression trong mọi trường hợp
- Khả năng dự đoán của 2 mô hình tương đối chính xác ở các mức giá khách sạn từ vừa đến thấp (<2 triệu đồng) và giảm dần khi mức giá tăng. Nguyên nhân cho việc này là bởi vì tập dữ liệu huấn luyện có ít các khách sạn với mức giá cao cho mô hình học, dẫn đến mô hình không thể dự đoán chính xác cho các khách sạn cao cấp.
- Hầu hết các dự đoán ở mức giá cao đều thấp hơn so với mức giá thực tế. Điều này càng khẳng định kết luận trên của chúng ta là chính xác.

4.3.2. Kết quả bài toán Classification

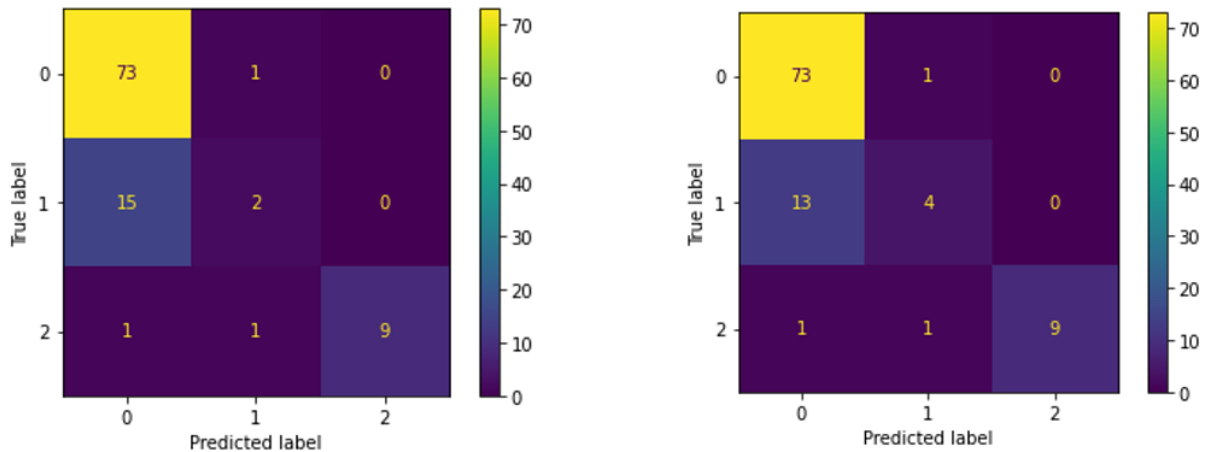
- Nhóm sử dụng 3 metrics : Accuracy và F1 score cùng với Confusion Matrix để đánh giá kết quả huấn luyện của mô hình

Bảng 8: Bảng đánh giá hiệu suất mô hình trên tập xác thực

Thuật toán	Accuracy	F1 score
Softmax Regression	0.89	0.79
Random Forest	0.89	0.80
Random Forest sau khi fine-tuned	0.90	0.81

Bảng 9: Bảng đánh giá hiệu suất mô hình trên tập kiểm thử

Thuật toán	Accuracy	F1 score
Softmax Regression	0.82	0.66
Random Forest	0.84	0.72



Hình 20: Ma trận nhầm lẫn của Softmax Regression (trái) và Random Forest (phải)

- Từ các kết quả trên, ta thu được các kết luận sau:

- Softmax Regression và Random Forest cho kết quả tương đương nhau trên tập xác thực nhưng Random Forest làm tốt hơn tương đối trên tập kiểm thử
- Phần lớn sự phân loại nhầm lẫn là từ hạng cao cấp về hạng thông thường. Điều này có thể gây ra bởi không có một ranh giới cụ thể giữa 2 hạng phòng này trong tập huấn luyện, khiến cho mô hình không thể phân loại tốt 2 hạng phòng này.

5. Kết luận

5.1. Kết quả thu được

- Từ kết quả EDA, ta có thể thấy được các biến Size, Distance to Beach và Distance to airport có ảnh hưởng cao đến mức giá phòng. Ngoài ra 2 biến Pool và Bar cũng có tác động đến mức giá của các phòng cùng kích thước.

- Hầu hết các khách sạn và đa số các khách sạn đắt tiền được xây gần biển hơn gần sân bay. Các chủ đầu tư nên cân nhắc xây dựng khách sạn của mình ở gần biển nếu muốn nâng cao giá phòng.

- Mô hình Random Forest trong bài toán hồi quy cho ra kết quả tốt hơn Linear Regression. Tuy nhiên cả 2 mô hình đều có hiệu suất giảm dần khi dự đoán các mức giá cao. Nguyên nhân chính của việc này là số mẫu dữ liệu ở mức giá này không đủ nhiều cho việc huấn luyện mô hình.
- Đối với bài toán phân loại, cả 2 mô hình cho ra kết quả tương đương nhau. Cả 2 đều không thể phân loại tốt giữa hạng phòng thường và hạng cao cấp. Nguyên nhân khả năng cao bởi vì sự không rõ ràng giữa 2 hạng phòng này ở các đặc trưng được chọn.

5.2. Hướng phát triển

- Thu thập thêm các mẫu dữ liệu ở các phân khúc cao cấp.
- Nghiên cứu thêm về các đặc trưng mới để phân loại hạng phòng thường và hạng cao cấp. Ngoài ra có thể nghiên cứu thêm về việc phân cấp các hạng phòng cụ thể hơn để mô hình có thể học tốt hơn.

6. Tài liệu tham khảo

- [1] H. T. Vu, «Bài 3: Linear Regression,» 28 12 2016. [En ligne]. Available: <https://machinelearningcoban.com/2016/12/28/linearregression/>.
- [2] H. T. Vu, «Bài 13: Softmax Regression,» 17 2 2017. [En ligne]. Available: <https://machinelearningcoban.com/2017/02/17/softmax/>.
- [3] H. T. Vu, «Bài 11: Giới thiệu về Feature Engineering,» 6 2 2017. [En ligne]. Available: <https://machinelearningcoban.com/general/2017/02/06/featureengineering/>.
- [4] T. Nguyen, «Random Forest Algorithm,» [En ligne]. Available: https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html.