ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

SỬ DỤNG CÁC KỸ THUẬT PHÂN TÍCH KẾT HỢP VỚI PHƯƠNG PHÁP HỌC MÁY DỰ ĐOÁN SỐ LƯỢNG XE ĐẠP CHO THUÊ MỖI GIỜ TẠI SEOUL

STT	Họ tên	MSSV
1	Trần Nguyễn Anh Khoa	18520938
2	Nguyễn Hoàng Huy	18520842

TP. HỒ CHÍ MINH – 12/2020

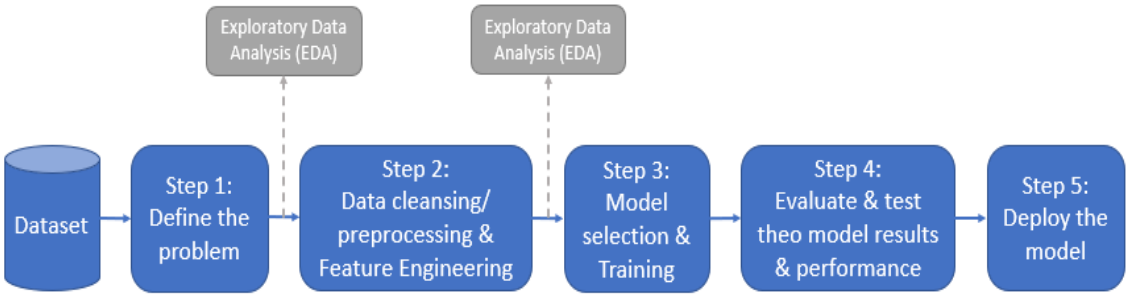
1. GIỚI THIỆU

Hiện nay xe đạp cho thuê đã được triển khai ở nhiều thành phố đô thị để nâng cao sự thoải mái khi di chuyển. Điều quan trọng là việc cung cấp xe đạp cho thuê phải tiếp cận được với công chúng vào đúng thời điểm vì nó sẽ giảm bớt thời gian chờ đợi. Do đó, việc đáp ứng cho các địa điểm thuê xe trong thành phố có một nguồn cung cấp xe đạp cho thuê ổn định trở thành một mối quan tâm lớn. Ở đề tài này, chúng tôi sẽ vận dụng phương pháp học máy (machine learning) để dự đoán số xe đạp được cho thuê vào mỗi giờ trong ngày để có thể chuẩn bị nguồn cung cấp xe đạp cho thuê ổn định.

Với bộ dữ liệu có được, chúng tôi sẽ tiến hành tiền xử lý, phân tích thăm dò tìm ra những đặc trưng quan trọng nhất để phục vụ cho việc huấn luyện mô hình. Các mô hình được sử dụng trong đề tài này là Polynomial Regression, Random Forest Regression, sau khi huấn luyện chúng tôi sẽ tiến hành kiểm tra kết quả dự đoán, so sánh và đánh giá hai mô hình. Kết quả khi đánh giá bằng phương pháp cross validation, mô hình tốt nhất của Polynomial Regression cho kết quả R2 = 0.8358, mô hình tốt nhất của Random Forest Regression cho kết quả R2 = 0.8597. Ở đề tài này, mô hình Random Forest Regression sẽ được sử dụng để tiến hành dự đoán.

2. NỘI DUNG

Trong đề tài này, phần nội dung sẽ được thực hiện bám sát theo quy trình ở hình bên dưới:



Hình 1. Quy trình PTDL.

2.1. Giới thiệu bộ dữ liệu

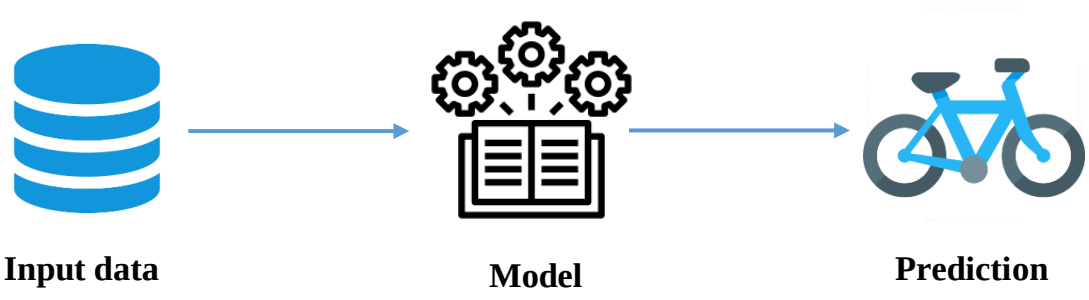
- Tên bộ dữ liệu: Seoul Bike Sharing Demand Dataset.
- Nguồn dữ liệu: UCI Machine Learning Repository: Seoul Bike Sharing Demand Data Set.
- Ý nghĩa bộ dữ liệu: Tập dữ liệu chứa số lượng xe đạp công cộng được thuê mỗi giờ trong hệ thống cho thuê xe đạp Seoul với dữ liệu thời tiết và thông tin ngày lễ tương ứng.
- Số biến (feature): 14.
- Số dòng (sample): 8760.
- Dữ liệu thiếu: không có.

- Mô tả cho mỗi biến (feature):

STT	Tên biến	Ý nghĩa	Miền giá trị	Kiểu dữ liệu
1	Date	Ngày/ tháng/ năm	01/12/2017 – 30/11/2018	Object
2	Rented bike count	Số xe đạp được thuê theo từng khung giờ	[0, 3556]	Integer
3	Hour	Giờ trong ngày	[0, 23]	Integer
4	Temperature	Nhiệt độ trong từng khung giờ (độ C)	[-17.8, 39.4]	Float
5	Humidity	Độ ẩm trong từng khung giờ (%)	[0, 98]	Integer
6	Windspeed	Tốc độ gió trong từng khung giờ (m/s)	[0, 7.4]	Float
7	Visibility	Tầm nhìn trong bán kính 10m	[27, 2000]	Integer
8	Dew point temperature	Nhiệt độ điểm sương (độ C)	[-30.6, 27.2]	Float
9	Solar radiation	Độ bức xạ mặt trời (MJ/m2)	[0, 3.52]	Float
10	Rainfall	Độ dày mưa rơi (mm)	[0, 35]	Float
11	Snowfall	Độ dày tuyết rơi (cm)	[0, 8.8]	Float
12	Seasons	Mùa trong năm	{Spring, Summer, Autumn, Winter}	Object
13	Holiday	Ngày lễ	{Holiday, No holiday}	Object
14	Functioning day	Ngày mà dịch vụ cho thuê xe hoạt động hay không	{Yes, No}	Object

2.2. Xác định bài toán

Dựa vào dữ liệu thời tiết và thông tin về thời gian, dự đoán số lượng xe đạp được thuê mỗi giờ để có thể cung cấp nguồn xe đạp cho thuê ổn định. Từ bộ dữ liệu có sẵn, chúng tôi sẽ tiến hành tiền xử lý, phân tích dữ liệu, chọn lọc ra các biến có ảnh hưởng nhất đến biến phụ thuộc **Rented bike count** để xây dựng mô hình, các mô hình được dùng để huấn luyện với dữ liệu là Polynomial Regression và Random Forest Regression, sau đó sẽ so sánh hiệu suất giữa hai mô hình và chọn ra mô hình tốt nhất để giải quyết bài toán.

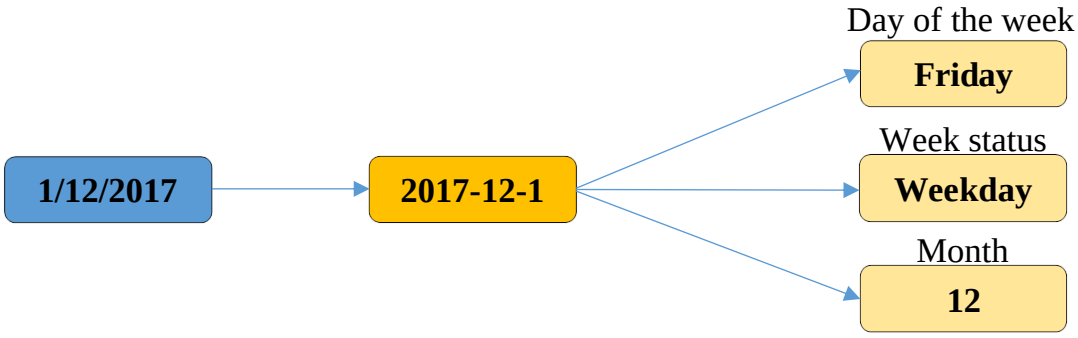


Hình 2. Bài toán đặt ra.

2.3. Tiền xử lý dữ liệu, phân tích thăm dò (EDA) và feature engineering

2.3.1. Tiền xử lý dữ liệu

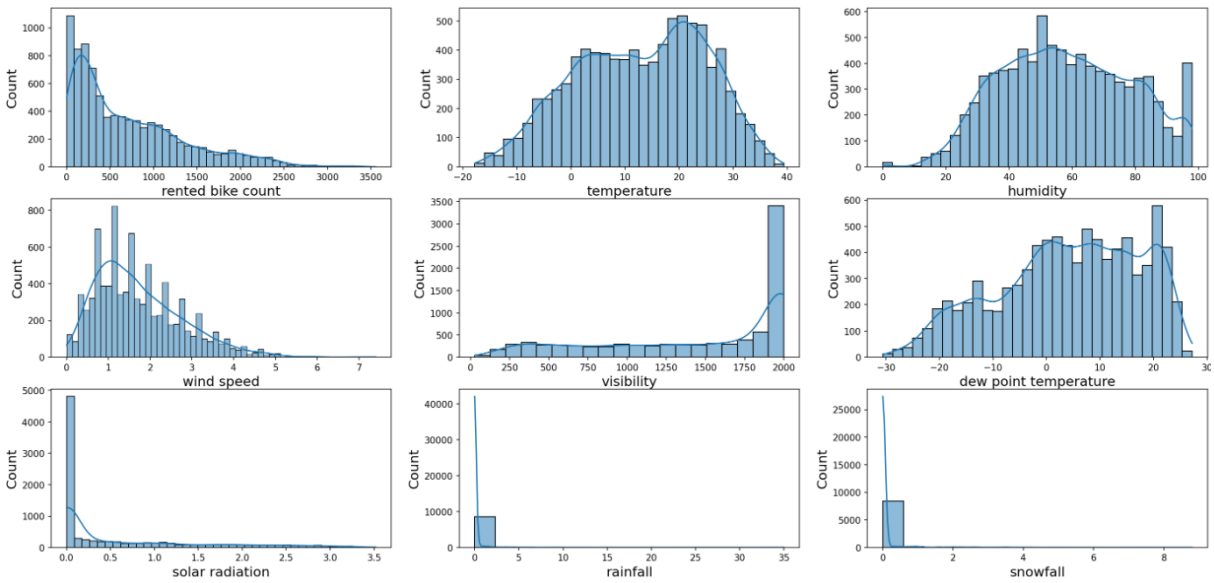
Ở biến **Date**, dữ liệu ngày tháng năm bên trong sẽ được chuyển đổi từ định dạng **dd/mm/yyyy** sang **yyyy-mm-dd** để có thể dễ dàng truy xuất và từ biến **Date** chúng tôi sẽ xử lý thành các biến **Day of the week (ngày trong tuần)**, **Week status (trạng thái của ngày trong tuần)**, **Month (tháng)** để phục vụ cho việc phân tích.



Hình 3. Xử lý dữ liệu biến Date.

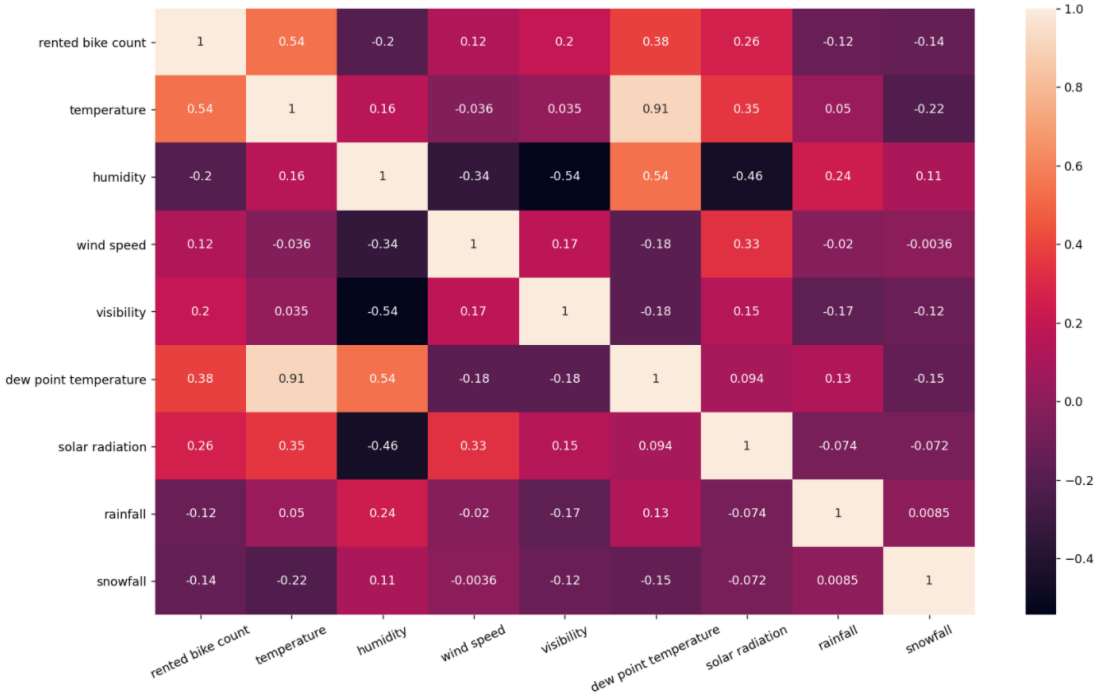
2.3.2. Phân tích thăm dò (EDA) và feature engineering

Sau bước tiền xử lý dữ liệu, tiến hành phân tích thăm dò là bước tiếp theo cần thực hiện để từ đó có thể chọn ra được những biến tốt nhất, có ảnh hưởng nhất để phát triển mô hình dự đoán. Đầu tiên chúng tôi sẽ thực hiện việc phân tích trên các biến số (numerical variables), biểu đồ **histogram** sẽ được dùng để quan sát sự phân phối của dữ liệu.



Hình 4. Biểu đồ histogram.

Quan sát hình 4, các biến **Temperature, Humidity, Dew point temperature** có dữ liệu phân phối khá cân đối, đây có thể là các biến quan trọng để phát triển mô hình, các biến **Windspeed, Solar radiation, Rainfall, Snowfall** có dữ liệu phân phối lệch về bên phải (right skewed), biến **Visibility** có dữ liệu phân phối lệch về bên trái (left skewed). Tiếp theo sử dụng **correlation matrix** để thể hiện sự tương quan của các biến độc lập lên biến phụ thuộc **Rented bike count** và tiến hành lựa chọn biến phục vụ cho việc phát triển mô hình**.**

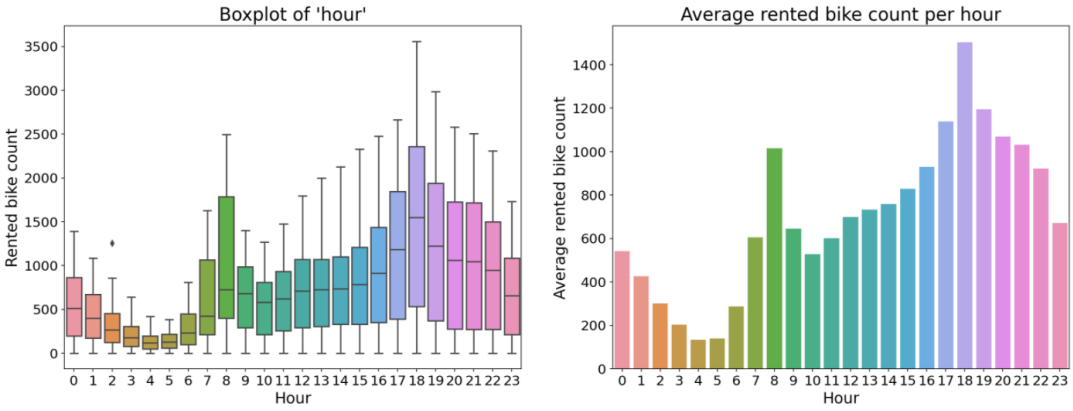


Hình 5. Correlation matrix.

Ở hình 5, độ tương quan đã tính toán giữa các biến độc lập so với biến **Rented bike count** đã được kiểm chứng là có độ tin cậy cao. Tiến hành chọn các biến số có độ tương quan (Pearson Coef) với **Rented bike count** từ [-0.2; -1) và [0.2; 1) để phát triển

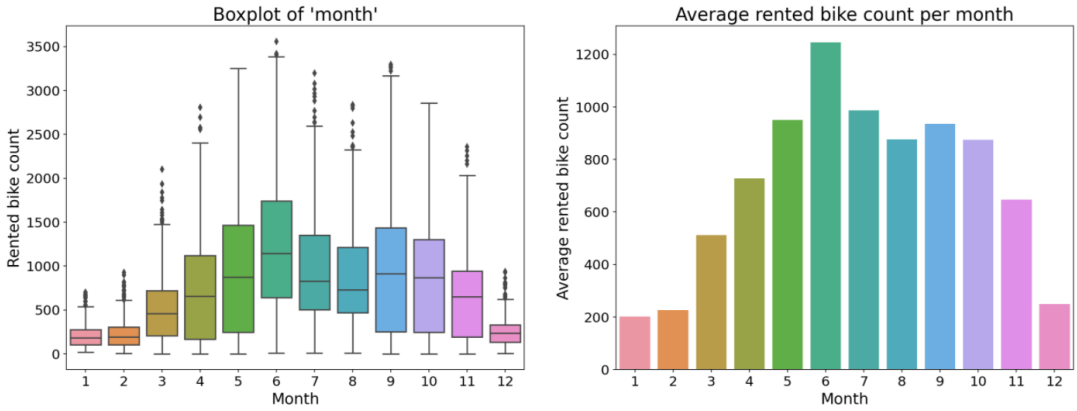
mô hình. Các biến được chọn là **Temperature**, **Humidity**, **Visibility**, **Dew point temperature**, **Solar radiation**.

Tiếp theo, việc lựa chọn các biến phân loại thích hợp để phát triển mô hình sẽ được thực hiện, **boxplot (biểu đồ hộp)** và **barplot (biểu đồ cột)** sẽ được dùng để phân tích các biến phân loại. Đầu tiên là biến **Hour**.



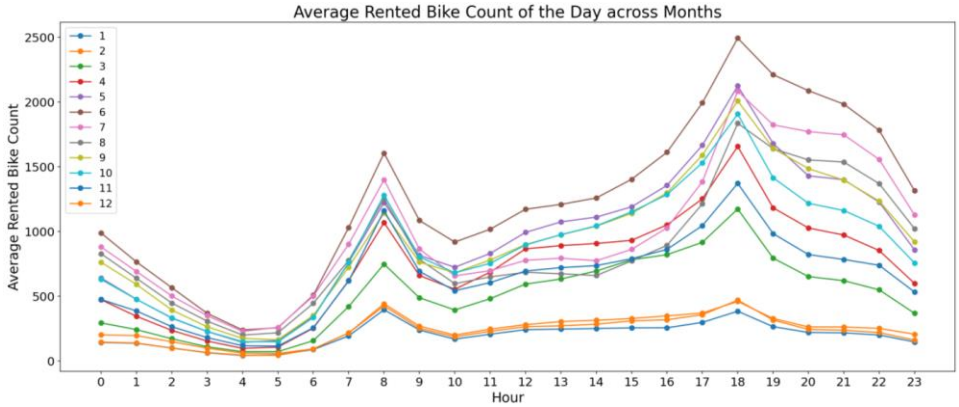
Hình 6. Boxplot và Barplot của biến Hour.

Quan sát hình 6, ở **Boxplot of 'hour'** gần như không xuất hiện các giá trị ngoại lệ, kết hợp quan sát với **Average rented bike count per hour** cho thấy ở từng khung giờ khác nhau sẽ có lượng xe đạp trung bình được cho thuê là khác nhau. Vào buổi sáng, số lượng xe cho thuê cao vào khoảng thời gian từ 7h-9h. Vào buổi trưa, số lượng xe được cho thuê duy trì ở một mức ổn định và có xu hướng tăng nhẹ. Vào chiều tối, số lượng xe cho thuê cao từ 17h-19h. Số lượng xe đạp cho thuê bắt đầu giảm từ 20h-23h và giảm mạnh từ 0h-4h. Qua đó, có thể thấy được rằng các giờ khác nhau sẽ có số lượng xe thuê khác nhau, giờ sẽ ảnh hưởng lên số lượng xe được cho thuê. Chúng tôi sẽ chọn biến **Hour**. Tiếp theo sẽ đến biến **Month**.



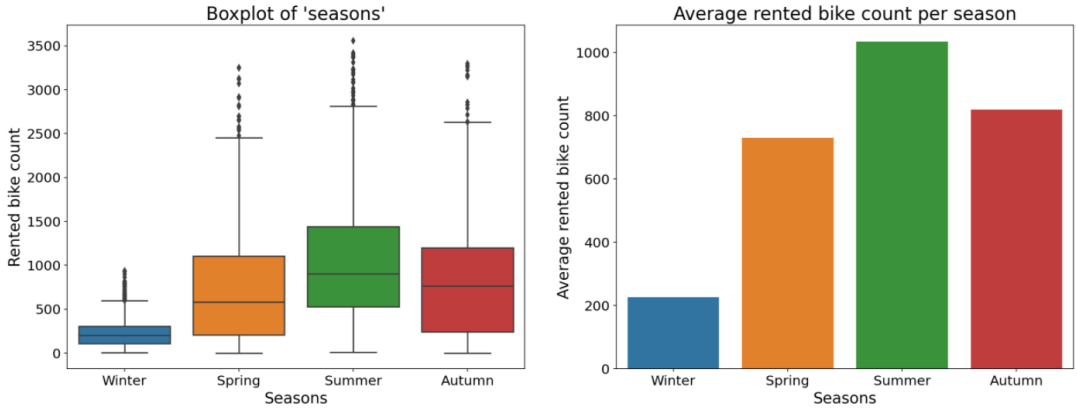
Hình 7. Boxplot và Barplot của biến Month.

Ở hình 7, quan sát **Boxplot of 'month'** chúng tôi thấy rằng hầu hết các tháng đều có giá trị ngoại lệ và kết hợp quan sát với **Average rented bike count per month** cho thấy các tháng khác nhau mang những đặc trưng thời tiết khác nhau cũng sẽ ảnh hưởng đến lượng xe đạp cho thuê, cụ thể là với tháng 12, 1, 2 (mùa đông ở Hàn Quốc) nhiệt độ hạ xuống thấp vì thế cư dân hạn chế đi ra ngoài nên số lượng xe đạp được thuê cũng giảm, các tháng còn lại thời tiết ôn hòa, ấm hơn nên có số lượng xe thuê cao hơn và nhu cầu thuê cao nhất vào giai đoạn tháng 6, 7, 8 (mùa hè ở Hàn Quốc, nhiệt độ cao hơn).



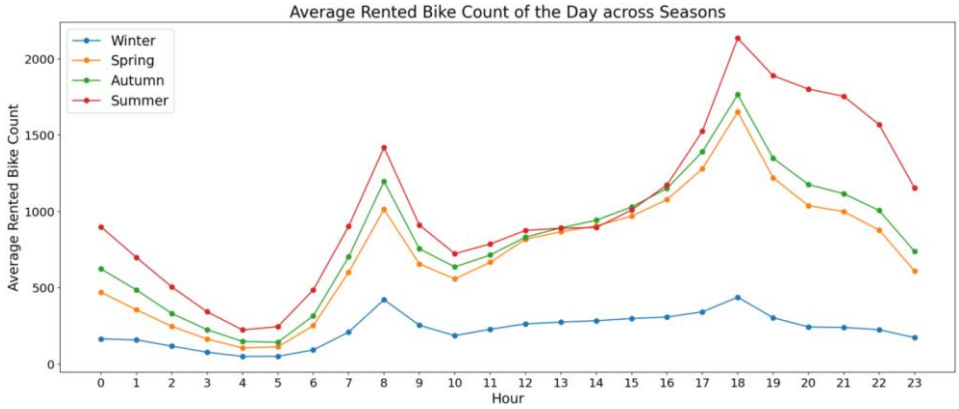
Hình 8. Lineplot thể hiện số xe đạp được thuê trung bình trong ngày qua các tháng.

Hình 8 thể hiện mối quan hệ giữa biến **Hour** và **Month**, khung giờ 7h-9h và 17h-19h có số lượng xe được thuê cao ở mỗi tháng vì đây là khoảng thời gian cao điểm và các tháng khác nhau có trung bình số lượng xe đạp cho thuê khác nhau. Qua đó cũng thấy được rằng biến **Month** có ảnh hưởng đến biến mục tiêu **Rented bike count**, chúng tôi sẽ chọn biến **Month** để phát triển mô hình. Tiếp theo là biến **Seasons**.



Hình 9. Barplot và Boxplot của biến Seasons.

Ở hình 9, các mùa đều xuất hiện giá trị ngoại lệ khi quan sát **Boxplot of 'seasons'**, khi kết hợp quan sát với **Average rented bike count per season** thì thấy được rằng chênh lệch về số lượng xe đạp cho thuê giữa các mùa thể hiện rõ ràng nhất ở mùa đông (Winter) và mùa hè (Summer). Còn mùa xuân (Spring), mùa thu (Autumn) không cho thấy sự chênh lệch ở số xe được cho thuê một cách rõ rệt.



Hình 10. Lineplot thể hiện số xe đạp được thuê trung bình trong ngày qua các mùa.

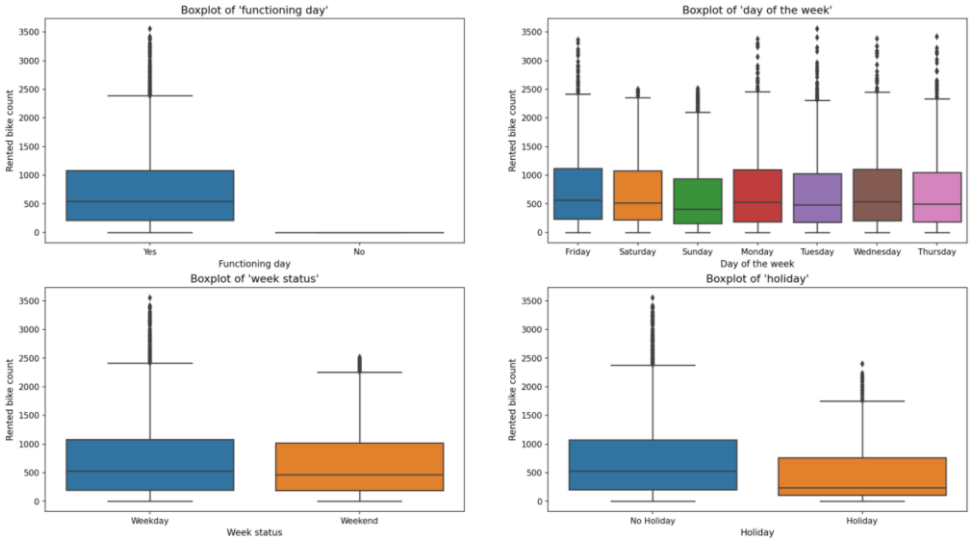
Hình 10 cho thấy mối liên hệ chặt chẽ giữa những người sử dụng xe đạp cho thuê và các yếu tố thời tiết kết hợp với ngày/giờ. Khung giờ từ 7h-9h, 17h-19h có số lượng xe thuê cao nhất ở mỗi mùa vì đây là khoảng thời gian cao điểm. Các mùa khác nhau với những đặc điểm thời tiết khác nhau dẫn đến số lượng xe cho thuê khác nhau, cụ thể mùa đông có số lượng xe cho thuê trung bình thấp nhất vì ở thời điểm này nhiệt độ thời tiết hạ xuống rất thấp, dân cư sẽ hạn chế ra ngoài. Ngược lại vào mùa hè, số lượng xe thuê trung bình cao nhất trong các mùa vì vào mùa hè nhiệt độ sẽ ấm nhất trong năm, nhu cầu thuê xe đạp cao và ổn định vì không gặp trở ngại về thời tiết. Mùa xuân và mùa thu có số lượng xe được thuê trung bình gần bằng nhau. Chúng tôi sẽ tiến hành phân tích ANOVA để xem sự tương quan (correlation) giữa các nhóm trong biến **Seasons** so với biến mục tiêu **Rented bike count** và sẽ giữ lại nhóm có correlation cao nhất để mã hóa (encode) chúng thành biến phục vụ cho việc phát triển mô hình.

	Values	F-test	P-value	Certainly
0	[Summer, Winter]	2832.243777	0.000000e+00	Strong
1	[Autumn, Winter]	1708.068675	3.805575e-315	Strong
2	[Spring, Winter]	1346.099529	4.102410e-257	Strong
3	[Summer, Autumn, Winter]	1234.655126	0.000000e+00	Strong
4	[Spring, Summer, Winter]	1223.899098	0.000000e+00	Strong
5	[Spring, Autumn, Winter]	798.143534	6.338928e-311	Strong
6	[Spring, Summer, Autumn, Winter]	776.467815	0.000000e+00	Strong
7	[Spring, Summer]	236.592354	4.771203e-52	Strong
8	[Spring, Summer, Autumn]	125.620382	2.864039e-54	Strong
9	[Summer, Autumn]	112.156453	6.762943e-26	Strong
10	[Spring, Autumn]	21.748938	3.199949e-06	Strong

Hình 11. Kết quả phân tích ANOVA giữa các nhóm giá trị trong biến Seasons.

Quan sát hình 11, nhóm [Summer,Winter] có giá trị F-test lớn nhất do đó nhóm [Summer,Winter] có correlation mạnh nhất với biến mục tiêu **Rented bike count**. Chúng tôi sẽ chọn nhóm giá trị này đại diện cho biến **Seasons** để phát triển mô hình.

Tiếp theo chúng tôi sẽ dùng **boxplot** để phân tích các biến phân loại **Functioning day, Day of the week, Week status, Holiday** và ra quyết định.



Hình 12. Boxplot của các biến Functioning day, Day of the week, Week status, Holiday.

Quan sát ở hình 12, dựa vào boxplot cho thấy các biến **Day of the week, Week status, Holiday** có sự phân hóa về số xe được thuê ở từng giá trị gần như tương đồng nhau, vì thế chúng tôi sẽ không chọn các biến này để phát triển mô hình. Ở biến **Functioning day**, có giá trị *Yes* thì sẽ có xe thuê, còn nếu có giá trị *No* thì sẽ không có xe thuê vì đây là ngày mà dịch vụ cho thuê xe không hoạt động; hai giá trị này tạo ra sự phân hóa rõ rệt ở số lượng xe được thuê. Do đó, biến **Functioning day** có ảnh hưởng đến biến mục tiêu **Rented bike count**, chúng tôi sẽ chọn biến này để phát triển mô hình.

Các biến phân loại được lựa chọn để phát triển mô hình là **Hour**, **Month**, **Seasons**, **Functioning day**. Sau đó chúng tôi sẽ mã hóa (encode) các biến có giá trị chữ sang dạng số để các mô hình phân tích có thể sử dụng được, **Functioning day** với giá trị là *Yes/No* sẽ được mã hóa thành *1/0* (hình 13).

functioning day
Yes
No

→

functioning_day
1
0

Hình 13. Encode biến Functioning day.

Sau đó, **Seasons** với 4 giá trị *Spring, Summer, Autumn, Winter* được tạo thành 4 biến theo phương pháp **get_dummies** và 2 biến **Season_Summer, Season_Winter** sẽ được chọn để phát triển mô hình (hình 14) dựa trên kết quả phân tích ANOVA (hình 11).

seasons
Winter
Winter
Spring
Spring

→

Season_Autumn	Season_Spring	Season_Summer	Season_Winter
0	0	0	1
0	0	0	1
0	1	0	0
0	1	0	0

Hình 14. Encode biến Seasons.

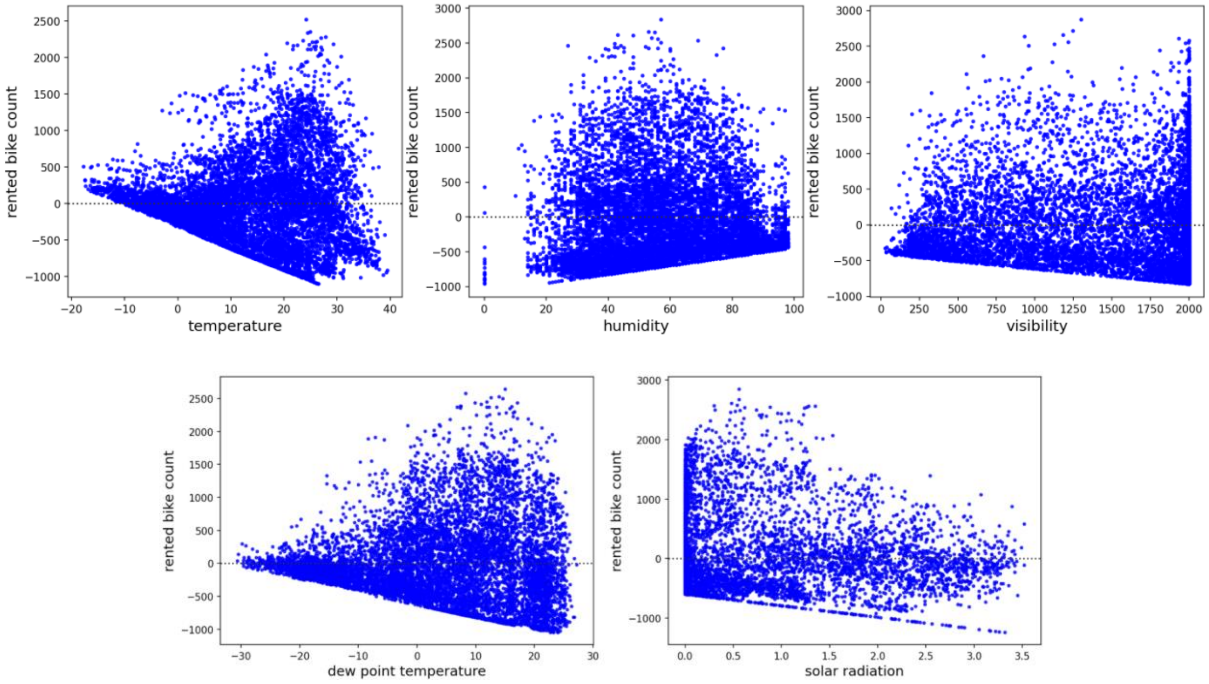
Qua bước phân tích thăm dò, feature engineering và mã hóa các biến phân loại, các biến được chọn là **Temperature, Humidity, Visibility, Dew point temperature, Solar radiation, Hour, Month, Season_Summer, Season_Winter, Functioning day**. Sau đó, chúng tôi sẽ tiến hành tạo ra bộ dữ liệu từ các biến này để phục vụ cho việc phát triển mô hình.

	hour	month	temperature	humidity	visibility	dew point temperature	solar radiation	functioning_day	Season_Summer	Season_Winter	rented bike count
0	0	12	-5.2	37	2000	-17.6	0.0	1	0	1	254
1	1	12	-5.5	38	2000	-17.6	0.0	1	0	1	204
2	2	12	-6.0	39	2000	-17.7	0.0	1	0	1	173
3	3	12	-6.2	40	2000	-17.6	0.0	1	0	1	107
4	4	12	-6.0	36	2000	-18.6	0.0	1	0	1	78
...	...	...	...	...	...	...	...	...	...	...	...
8755	19	11	4.2	34	1894	-10.3	0.0	1	0	0	1003
8756	20	11	3.4	37	2000	-9.9	0.0	1	0	0	764
8757	21	11	2.6	39	1968	-9.9	0.0	1	0	0	694
8758	22	11	2.1	41	1859	-9.8	0.0	1	0	0	712
8759	23	11	1.9	43	1909	-9.3	0.0	1	0	0	584

Hình 15. Bộ dữ liệu phục vụ cho việc phát triển mô hình.

2.4. Phát triển mô hình, so sánh hiệu suất và đánh giá

2.4.1. Phát triển mô hình



Hình 16. Residual plot của các biến số được chọn để phát triển mô hình.

Quan sát ở hình 16, chúng tôi thấy các thuộc tính có điểm dữ liệu không ngẫu nhiên trải đều quanh trục x nên các thuộc tính này không phù hợp để phát triển mô hình hồi quy tuyến tính. Do đó, mô hình Polynomial Regression và Random Forest Regression sẽ được sử dụng để giải quyết bài toán đặt ra. Chúng tôi tiến hành tạo ra danh sách các mô hình của 2 thuật toán, danh sách này sẽ thể hiện các thông tin về mô hình, các biến dùng để huấn luyện mô hình, các độ đo dùng để đánh giá mô hình: R2 Score (train, test, 4-fold cross validation, 5-fold cross validation), RMSE và độ lệch chuẩn của kết quả R2 Score khi đánh giá bằng kỹ thuật cross validation. Với 10 biến độc lập dùng để phát triển mô hình, sẽ có 1023 trường hợp tập con từ các biến độc lập này để huấn luyện mô hình, ở mô hình Polynomial Regression sẽ được vét cạn đến *bậc 4 (degree=4)*, mô hình Random Forest Regression sẽ được vét cạn với các tham số *n_estimators = 15 (số cây*

quyết định) kết hợp lần lượt với *max_depth = {5, 10, 15} (độ sâu của cây)* ứng với mỗi tập con đặc trưng khác nhau**.** Từ danh sách các mô hình vừa tạo ra, chúng tôi tiến hành tìm ra các mô hình tốt nhất của Polynomial Regression (hình 17) và Random Forest Regression (hình 18) dựa vào kết quả mean R2 Score (5-fold cross validation).

	Model	Feature details	RMSE	R^2 train	R^2 test	4 fold - mean	4 fold - std	5 fold - mean	5 fold - std	note:
0	Polynomial Regression	['hour', 'month', 'temperature', 'humidity', '...	263.759314	0.841843	0.833775	0.833002	0.005435	0.832517	0.005907	test_size = 0.2, Degree = 4, num feature = 7
1	Polynomial Regression	['hour', 'month', 'temperature', 'humidity', '...	264.731902	0.840521	0.832547	0.833299	0.004761	0.833107	0.004325	test_size = 0.2, Degree = 4, num feature = 7
2	Polynomial Regression	['hour', 'month', 'temperature', 'dew point te...	265.617143	0.837044	0.831425	0.830582	0.006630	0.830137	0.005247	test_size = 0.2, Degree = 4, num feature = 7
3	Polynomial Regression	['hour', 'month', 'humidity', 'dew point tempe...	265.549046	0.840049	0.831511	0.832148	0.005262	0.831770	0.004847	test_size = 0.2, Degree = 4, num feature = 7
4	Polynomial Regression	['hour', 'month', 'temperature', 'humidity', '...	261.672947	0.846485	0.836394	0.835846	0.004953	0.835867	0.005276	test_size = 0.2, Degree = 4, num feature = 8
5	Polynomial Regression	['hour', 'month', 'temperature', 'dew point te...	262.188425	0.843334	0.835749	0.833447	0.006761	0.833260	0.006076	test_size = 0.2, Degree = 4, num feature = 8
6	Polynomial Regression	['hour', 'month', 'humidity', 'dew point tempe...	265.969034	0.845894	0.830978	0.833898	0.005951	0.833454	0.006093	test_size = 0.2, Degree = 4, num feature = 8

Hình 17. Danh sách các mô hình Polynomial Regression tốt nhất.

Quan sát hình 17, chúng tôi thu được 7 mô hình Polynomial Regression tốt nhất với kết quả mean R2 Score (5-fold cross validation) đạt trên 0.83, các mô hình tốt nhất này được huấn luyện với trường hợp 7, 8 biến và đều có số bậc là 4 *(degree=4)*.

	Model	Feature details	RMSE	R^2 train	R^2 test	4 fold - mean	4 fold - std	5 fold - mean	5 fold - std	note
0	Random Forest Regressor	['hour', 'month', 'temperature', 'humidity', '...	249.890322	0.965598	0.850796	0.857173	0.005826	0.858712	0.005740	num ft = 7, n_estimators=15, max_depth = 15
1	Random Forest Regressor	['hour', 'month', 'temperature', 'humidity', '...	248.234882	0.967804	0.852766	0.858281	0.004844	0.859306	0.006527	num ft = 8, n_estimators=15, max_depth = 15
2	Random Forest Regressor	['hour', 'month', 'temperature', 'humidity', '...	249.222841	0.965574	0.851592	0.856133	0.005242	0.858880	0.006253	num ft = 8, n_estimators=15, max_depth = 15
3	Random Forest Regressor	['hour', 'month', 'temperature', 'humidity', '...	251.353968	0.965538	0.849043	0.857498	0.005787	0.859782	0.007892	num ft = 8, n_estimators=15, max_depth = 15
4	Random Forest Regressor	['hour', 'month', 'temperature', 'humidity', '...	249.490262	0.967113	0.851274	0.858179	0.003587	0.859194	0.006841	num ft = 9, n_estimators=15, max_depth = 15
5	Random Forest Regressor	['hour', 'month', 'temperature', 'humidity', '...	249.588479	0.965253	0.851156	0.856970	0.005233	0.859462	0.007243	num ft = 9, n_estimators=15, max_depth = 15
6	Random Forest Regressor	['hour', 'month', 'temperature', 'humidity', '...	248.176136	0.966540	0.852836	0.858933	0.002514	0.858967	0.007297	num ft = 9, n_estimators=15, max_depth = 15
7	Random Forest Regressor	['hour', 'month', 'temperature', 'humidity', '...	249.616708	0.967659	0.851123	0.858439	0.004052	0.858686	0.005833	num ft = 10, n_estimators=15, max_depth = 15

Hình 18. Danh sách các mô hình Random Forest Regression tốt nhất.

Quan sát hình 18, chúng tôi có được 8 mô hình Random Forest Regression tốt nhất với kết quả mean R2 Score (5-fold cross validation) đạt trên 0.85, các mô hình tốt nhất này được huấn luyện với trường hợp từ 7 đến 10 biến và có giá trị tham số *n_estimators, max_depth = 15.*

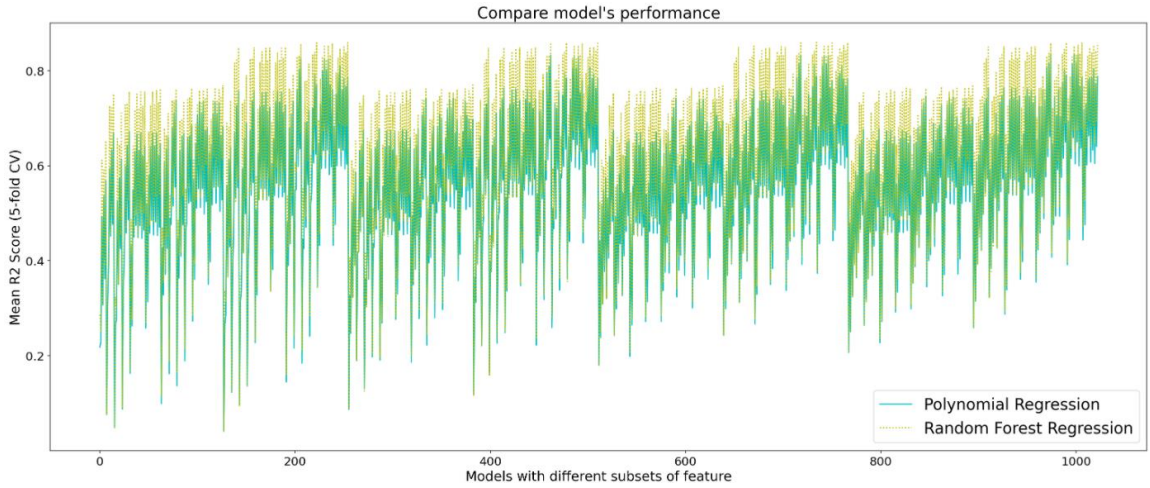
2.4.2. So sánh hiệu suất và đánh giá

Mô hình	R2 Score (5-fold CV)	RMSE	Biến
Polynomial Regression	0.835867	261.673	['hour', 'month', 'temperature', 'humidity', 'solar radiation', 'functioning_day', 'Season_Summer', 'Season_Winter']
Random Forest Regression	0.859782	251.353968	['hour', 'month', 'temperature', 'humidity', 'visibility', 'solar radiation', 'functioning_day', 'Season_Winter']

Hình 19. Mô hình tốt nhất của Polynomial Regression và Random Forest Regression.

Ở hình 19, có thể thấy rằng mô hình tốt nhất của Random Forest Regression cho kết quả cao hơn so với mô hình tốt nhất của Polynomial Regression, cả 2 mô hình này

đều được huấn luyện với 8 biến và các biến này khác nhau ở 2 mô hình. Tiếp theo, để có thể chọn ra mô hình dùng để triển khai và tiến hành dự đoán (deploy), chúng tôi sẽ tiến hành so sánh hiệu suất giữa Polynomial Regression với Random Forest Regression bằng mean R2 Score (5-fold cross validation) khi mô hình được huấn luyện ở từng trường hợp tập con đặc trưng khác nhau (ứng với mỗi tập con đặc trưng, mô hình Polynomial Regression và Random Forest Regression sẽ dùng mô hình cho kết quả tốt nhất để so sánh).

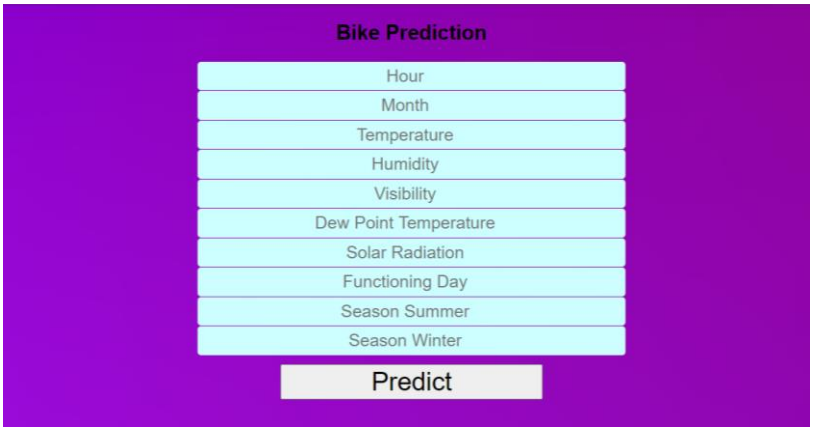


Hình 20. Biểu đồ so sánh hiệu suất giữa mô hình tốt nhất của Polynomial Regression và Random Forest Regression ứng với mỗi tập con đặc trưng khác nhau.

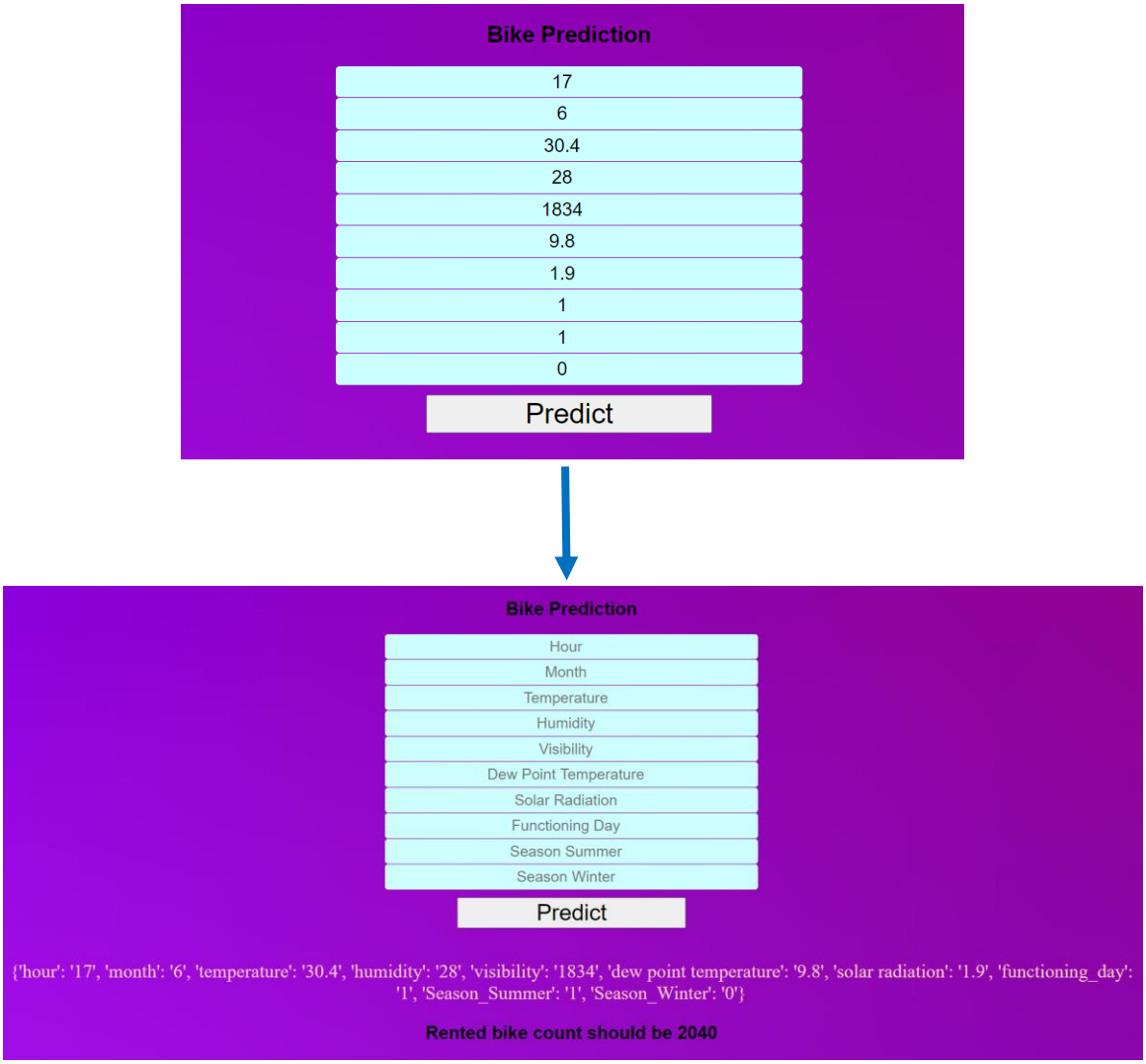
Quan sát hình 20, đường chấm màu vàng (Random Forest Regression) luôn nằm trên đường màu xanh ngọc (Polynomial Regression) ở hầu hết các trường hợp. Điều đó cho thấy rằng mô hình Random Forest Regression có hiệu suất tốt hơn so với Polynomial Regression. Chúng tôi sẽ chọn mô hình Random Forest Regression để triển khai mô hình và tiến hành dự đoán.

2.5. Triển khai mô hình

Mô hình Random Forest Regression sẽ được huấn luyện với toàn bộ dữ liệu để phục vụ cho việc triển khai và tiến hành dự đoán. Vận dụng các kiến thức về HTML, kết hợp cùng framework Flask trong python và với sự hỗ trợ của nền tảng Heroku (nền tảng đám mây cho phép triển khai ứng dụng), chúng tôi đã triển khai mô hình dự đoán thành công tại website với url: https://ds105-final-bike-model-deploy.herokuapp.com .



Hình 21. Giao diện của website dự đoán.



Hình 22. Dự đoán số xe được cho thuê với các giá trị đầu vào.

3. KẾT LUẬN

Với bộ dữ liệu Seoul Bike Sharing Demand, chúng tôi đã tiến hành tiền xử lý, phân tích thăm dò (EDA), feature engineering và mã hoá các biến phân loại có giá trị chữ để chọn ra được 10 biến tốt nhất, có ảnh hưởng nhất đến biến mục tiêu **Rented bike count**: **Hour, Month, Temperature, Humidity, Visibility, Dew point temperature, Solar radiation, Functioning_day, Season_Summer, Season_Winter**. Từ đó, chúng tôi sẽ tạo ra bộ dữ liệu phục vụ cho việc phát triển mô hình. Mô hình Polynomial Regression, Random Forest Regression được sử dụng để giải quyết bài toán này. Kết quả khi đánh giá bằng phương pháp cross validation (5-fold), mô hình tốt nhất của Polynomial Regression cho kết quả R2 = 0.8358, mô hình tốt nhất của Random Forest Regression cho kết quả R2 = 0.8597. Kết quả khi so sánh hiệu suất giữa Polynomial Regression với Random Forest Regression bằng mean R2 Score (5-fold cross validation) khi mô hình được huấn luyện ở từng trường hợp tập con đặc trưng khác nhau thì mô hình Random Forest Regression có hiệu suất tốt hơn so với Polynomial Regression ở hầu hết các trường hợp, do đó Random Forest Regression sẽ được lựa chọn để triển khai mô hình.

TÀI LIỆU THAM KHẢO

[1] Cross-validation: evaluating estimator performance. Link: 3.1. Cross-validation: evaluating estimator performance — scikit-learn 0.24.0 documentation (scikit-learn.org) (Ngày truy cập: 15/12/2020).

[2] Linear Regression. Link: sklearn.linear_model.LinearRegression — scikit-learn 0.24.0 documentation (scikit-learn.org) (Ngày truy cập: 14/12/2020).

[3] Polynomial Features. Link: sklearn.preprocessing.PolynomialFeatures — scikit-learn 0.24.0 documentation (scikit-learn.org) (Ngày truy cập: 14/12/2020).

[4] Random Forest Regression. Link: sklearn.ensemble.RandomForestRegressor — scikit-learn 0.24.0 documentation (scikit-learn.org) (Ngày truy cập: 14/12/2020).

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Trần Nguyễn Anh Khoa	Thực hiện phân tích dữ liệu, phát triển mô hình, viết báo cáo, soạn slide báo cáo cuối kỳ.
2	Nguyễn Hoàng Huy	Thực hiện triển khai mô hình, kiểm tra báo cáo, soạn slide báo cáo tiến độ.

PHỤ LỤC CODE

Source code phân tích dữ liệu, phát triển mô hình

- **URL:** anhkhoatrannguyen259/DS105_Final (github.com)

Source code triển khai mô hình (deploy model)

- **URL:** davione112/Deploy-model-bike-prediction (github.com)