TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN CUỐI KỲ

HỌC PHẦN: KHOA HỌC DỮ LIỆU

TÊN ĐỀ TÀI

DỰ ĐOÁN SỐ LƯỢNG SẢN PHẨM ĐÃ BÁN TẠI TIKI.VN

HỌ VÀ TÊN SINH VIÊN	LỚP HỌC PHẦN	ĐIỂM BẢO VỆ
Nguyễn Thanh Khải	19N10	

ĐÀ NẴNG, 06/2022

TÓM TẮT

Tiki là một trong những trang thương mại điện tử lớn nhất Việt Nam. Với chính sách gắt gao để hạn chế hàng nhái, hàng giả nên những thông tin sản phẩm trên tiki có thể tin tưởng được. Nếu có nhu cầu lấy dữ liệu sản phẩm thì tiki là một nguồn tham khảo đáng tin cậy. Nên để có thể chọn ra những sản phẩm chất lượng tốt, được nhiều người tin mua thì việc thu thập dữ liệu từ tiki cũng như thực hiện những đánh giá và tính toán độ chính xác dựa vào những thông tin của sản phẩm là cần thiết. Chính vì thế nên em quyết định chọn đề tài này với mục đích dự đoán những sản phẩm như thế nào thì sẽ được nhiều người tin mua hơn.

BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Tự đánh giá theo 3 mức (Đã hoàn thành/Chưa hoàn thành/Không triển khai)
Nguyễn Thanh Khải	- Thu thập dữ liệu từ tiki.vn - Trích xuất đặc trưng - Mô hình hóa dữ liệu	Đã hoàn thành

MỤC LỤC

Nội dung

MỤC LỤC HÌNH ẢNH

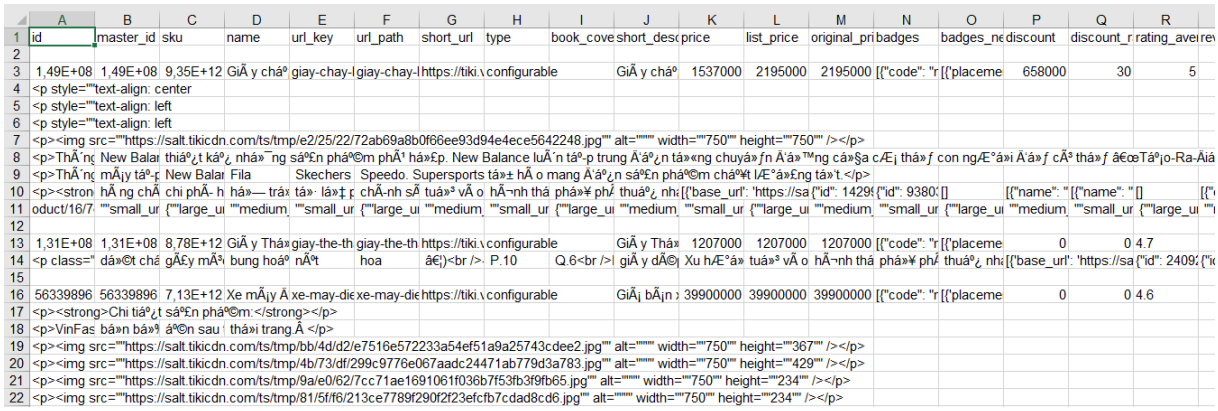
1. Giới thiệu

Từ dữ liệu đã thu thập được trên trang tiki.vn, em sẽ xây dựng các mô hình để dự đoán được số lượng sản phẩm đã được bán. Để làm được điều này cần phải làm sạch, chuẩn hóa dữ liệu đã thu thập sau đó tiến hành xây dựng mô hình để đưa ra đánh giá. Ở đây em chọn ra 2 mô hình để thực nghiệm là Linear Regression và Random Forest Regression. Từ 2 mô hình trên em sẽ so sánh kết quả thu được và chọn ra mô hình có kết quả tốt nhất để áp dụng.

2. Thu thập và mô tả dữ liệu

2.1. Thu thập dữ liệu

liệu được thu thập từ trang tiki.vn
Để crawl dữ liệu, cần chạy file crawl.py với input là các API "danh sách sản phẩm" đã tìm được trên trang tiki, đó là:
https://tiki.vn/api/v2/products?limit=48&include=advertisement&aggregations=2&trackity_id=7ac7bd0c-d274-5d0e-12e8-2c461a9e0820&q=s&page={}
https://tiki.vn/api/v2/products/{}
Và output là file product.csv (raw data) chứa các thông tin đã crawl về nhưng vẫn chưa qua xử lý.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	id	master_id	sku	name	url_key	url_path	short_url	type	book_cove	short_desc	price	list_price	original_pri	badges	badges_ne	discount	discount_r	rating_aver
2																		
3	1,49E+08	1,49E+08	9,35E+12	GiÃ y chá°	giay-chay-	giay-chay-l	https://tiki.v	configurable		GiÃ y chá°	1537000	2195000	2195000	[{"code": "r	[{"placeme	658000	30	5
4	<p style=""text-align: center																	
5	<p style=""text-align: left																	
6	<p style=""text-align: left																	
7																		
8	ThÃ´ng	New Balar	thiáº¿t káº¿ nhá»¯ng sáº£n phẩm phÃ¹ há»£p. New Balance luÃ´n táº·p trung Ã¡p dụng cá»«ng chuyá»ƒn Ã¡» ™ng cá»§a cÆ¡ thá»ƒ con ngÆ°á»i Ã¡p dá»¥ng cÃ³ thá»ƒ â€œTáº¡o-Ra-Ãiá															
9	ThÃ´ng	mÃ¡y táº·p	New Balar	Fila	Skechers	Speedo.	Supersports tá»± hÃ o mang Ã¡p dụng sáº£n phẩm chất lÆ°á»£ng tá»t.											
10	<stron	hÃ ng chÃ	chi phÃ- h	há»— trá»	tá»· lá»‡ p	chÃ-nh sÃ	tuá»³ vÃ o	hÃ¬nh thá	phá»¥ phÃ	thuáº¿ nhÃ	[{'base_url': 'https://sa	{"id": 1429	{"id": 9380	[]		[{"name": "	[{"name": "	[]	[{"
11	oduct/16/7	""small_ur	{""large_ui	""medium	""small_ur	{""large_ui	""medium	""small_ur	{""large_ui	""medium	""small_ur	{""large_ui	""medium	""small_ur	{""large_ui	""medium	""small_ur	{""large_ui	""
12																		
13	1,31E+08	1,31E+08	8,78E+12	GiÃ y Thá»	giay-the-th	giay-the-th	https://tiki.v	configurable		GiÃ y Thá»	1207000	1207000	1207000	[{"code": "r	[{"placeme	0	0	4.7
14	<p class="	dá»©t chá	gÃ£y mÃ³	bung hoáº	nÃºt	hoa	â€¦) .	P.10	Q.6 	giÃ y dÃ©p	Xu hÆ°á»	tuá»³ vÃ o	hÃ¬nh thá	phá»¥ phÃ	thuáº¿ nhÃ	[{'base_url': 'https://sa	{"id": 2409	{"
15																		
16	56339896	56339896	7,13E+12	Xe mÃ¡y Ã	xe-may-die	xe-may-die	https://tiki.v	configurable		GiÃ¡ bÃ¡n	39900000	39900000	39900000	[{"code": "r	[{"placeme	0	0	4.6
17	Chi tiáº¿t sáº£n pháº©m:																	
18	VinFas bá»n bá»‰ á®n sau thá»i trang.Â																	
19																		
20																		
21																		
22																		

Hình 1: raw data

Tiến hành chạy file clean.ipynb để làm sạch dữ liệu và lưu vào file output.csv (clean data) với 12 trường và 1000 mẫu dữ liệu.

id	name	type	price	original_price	discount	discount_r	rating	reviev	productset_gr	day_ago_c	all_time_quantity_sold
148803925	GiÃ y chá°¡y	configurable	1537000	2195000	658000	30	5.0	9	Thá»ƒ Thao -	185.0	15.0
131347928	GiÃ y Thá»ƒ	configurable	1207000	1207000	0	0	4.7	372	Thá»ƒ Thao -	238.0	752.0
56339896	Xe mÃ¡y Ä'iá	configurable	39900000	39900000	0	0	4.6	513	Ã" TÃ´ - Xe M	734.0	1175.0
118281790	GiÃ y Thá»ƒ	configurable	1086000	1207000	121000	10	4.6	38	Thá»ƒ Thao -	301.0	71.0
37468142	GiÃ y thá»ƒ t	configurable	699000	699000	0	0	1.0	1	Thá»ƒ Thao -	965.0	10.0
5204861	GiÃ y Thá»ƒ	configurable	885000	932000	47000	5	5.0	64	Thá»ƒ Thao -	1353.0	134.0
115453610	GiÃ y sneake	configurable	898000	1795000	897000	50	4.7	13	Thá»ƒ Thao -	323.0	21.0
147982904	GiÃ y Thá»ƒ	configurable	731000	731000	0	0	5.0	9	Thá»ƒ Thao -	186.0	22.0
148676118	GiÃ y Thá»ƒ	configurable	1167000	1229000	62000	5	5.0	3	Thá»ƒ Thao -	185.0	11.0
147982629	GiÃ y Thá»ƒ	configurable	731000	731000	0	0	5.0	11	Thá»ƒ Thao -	185.0	24.0
137971241	GiÃ y Ä'i bá»	configurable	795000	1590000	795000	50	5.0	3	Thá»ƒ Thao -	219.0	8.0
118281078	GiÃ y Thá»ƒ	configurable	1086000	1207000	121000	10	5.0	20	Thá»ƒ Thao -	301.0	37.0
176825442	Xe mÃ¡y Ä'iá	configurable	69900000	69900000	0	0	0.0	0	Ã" TÃ´ - Xe M	50.0	
96113354	GiÃ y chá°¡y	configurable	4860000	4860000	0	0	5.0	2	Thá»ƒ Thao -	399.0	4.0
118326573	GiÃ y Thá»ƒ	configurable	680000	716000	36000	5	5.0	12	Thá»ƒ Thao -	301.0	24.0
145379503	[á°¢nh thá°-t	configurable	199000	199000	0	0	0.0	0	Thá»ƒ Thao -	195.0	
145381828	[á°¢nh thá°-t	configurable	304000	304000	0	0	0.0	0	Thá»ƒ Thao -	195.0	
145377681	HÃ ng VNXK	configurable	304000	304000	0	0	0.0	0	Thá»ƒ Thao -	195.0	

Hình 2: clean data

2.2. Mô tả dữ liệu

Đọc dữ liệu từ file output.csv sẽ cho ra kết quả như hình dưới:

id	name	type	price	original_price	discount	discount_rate	rating_average	review_count	productset_group_name	day_ago_created	all_time_quantity_sold
148803925	Giày chạy bộ nam New Balance Classic - ML574	configurable	1537000	2195000	658000	30	5.0	9	Thể Thao - Dã Ngoại/Giày thể thao nam/Giày cha...	185.0	15.0
131347928	Giày Thể Thao Nam Biti's Hunter X Z MIDNIGHT B...	configurable	1207000	1207000	0	0	4.7	372	Thể Thao - Dã Ngoại/Giày thể thao nam/Giày cha...	238.0	752.0
56339896	Xe máy điện VinFast Klara S	configurable	39900000	39900000	0	0	4.6	513	Ô Tô - Xe Máy - Xe Đạp/Xe điện/Xe máy điện	734.0	1175.0
118281790	Giày Thể Thao Nam Biti's Hunter X Z Collection...	configurable	1086000	1207000	121000	10	4.6	38	Thể Thao - Dã Ngoại/Giày thể thao nam/Giày cha...	301.0	71.0

Hình 3: Thông tin 4 mẫu đầu tiên

Và được mô tả như sau:

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	1000 non-null	int64
1	id	1000 non-null	int64
2	name	1000 non-null	object
3	type	1000 non-null	object
4	price	1000 non-null	int64
5	original_price	1000 non-null	int64
6	discount	1000 non-null	int64
7	discount_rate	1000 non-null	int64
8	rating_average	1000 non-null	float64
9	review_count	1000 non-null	int64
10	productset_group_name	1000 non-null	object
11	day_ago_created	999 non-null	float64
12	all_time_quantity_sold	245 non-null	float64

Hình 4: Mô tả dữ liệu

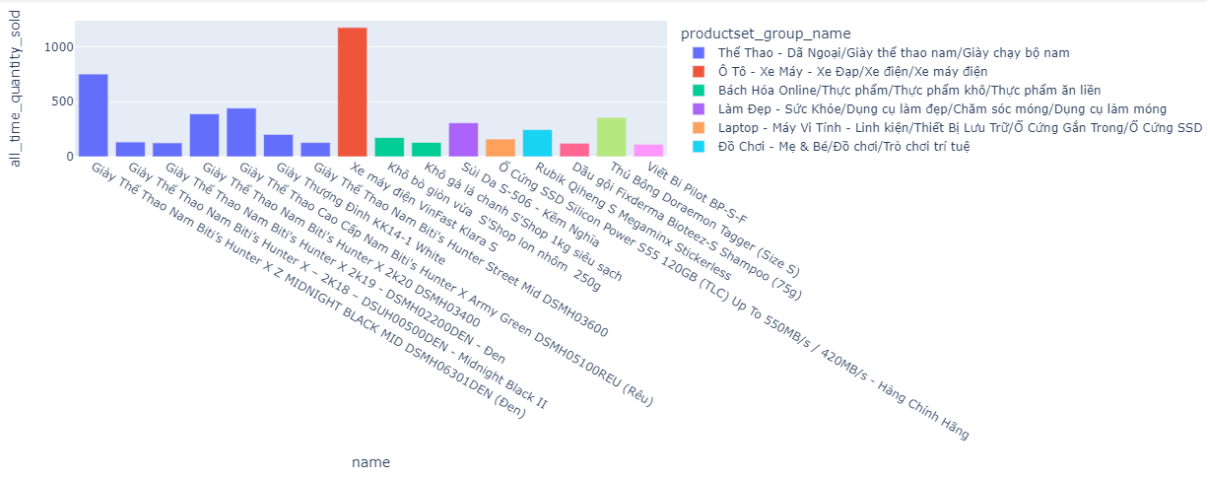
Theo đó, bộ dữ liệu sẽ có 1000 mẫu và mỗi mẫu có 12 đặc trưng. (1 đặc trưng Unnamed chứa số thứ tự)

Các đặc trưng có kiểu dữ liệu là số nguyên và số thực là: id, price, original_price, discount, discount_rate, rateing_average, review_count, day_ago_created, all_time_quantity_sold

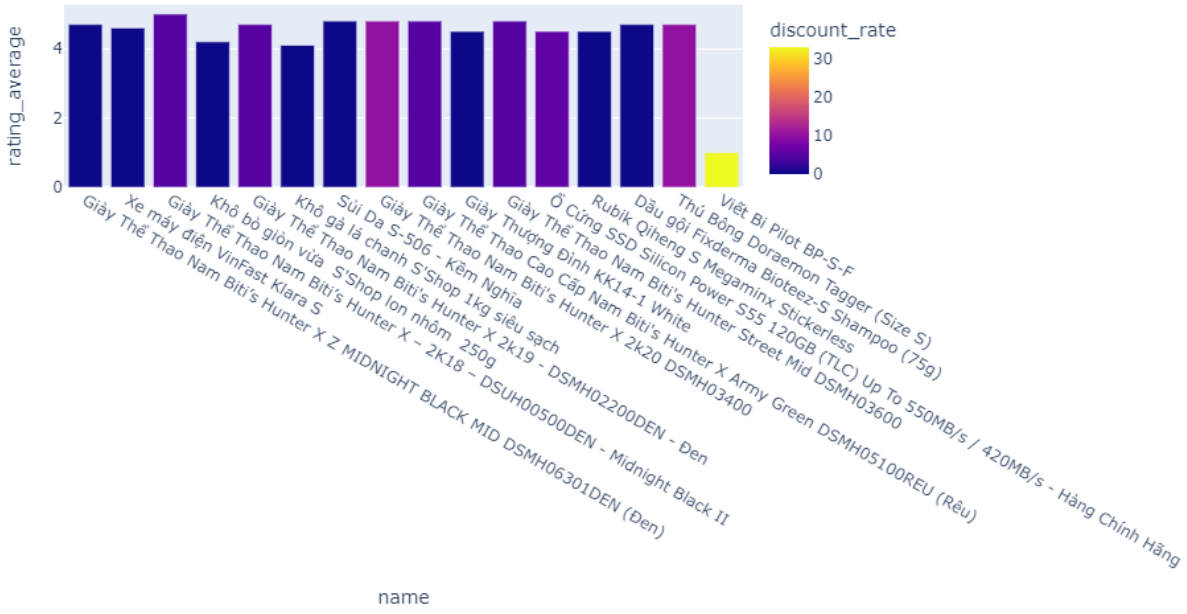
Các đặc trưng là có kiểu dữ liệu là chuỗi kí tự là: name, type, productset_group_name

Chỉ có 2 trường có dữ liệu trống là: day_ago_created (1 null) và all_time_quantity_sold (755 null). Trong đó trường all_time_quantity_sold có nhiều dữ liệu trống là bởi những sản phẩm này chưa được bán lần nào trên trang tiki nên thay vì để "0 đã bán" thì người ta không bổ sung dữ liệu vào trường này. Ở phần xử lý dữ liệu em sẽ thay thế các dữ liệu trống này bằng giá trị 0 cho đúng với thực tế.

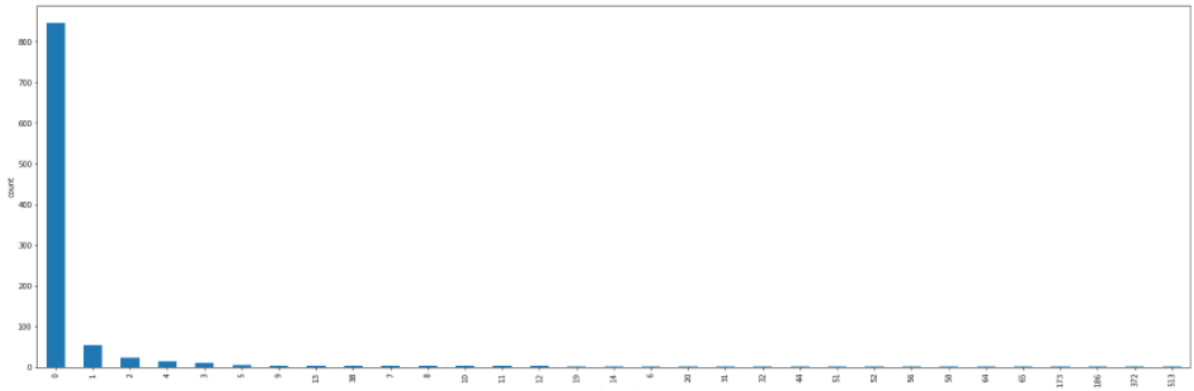
Các thống kê mô tả:



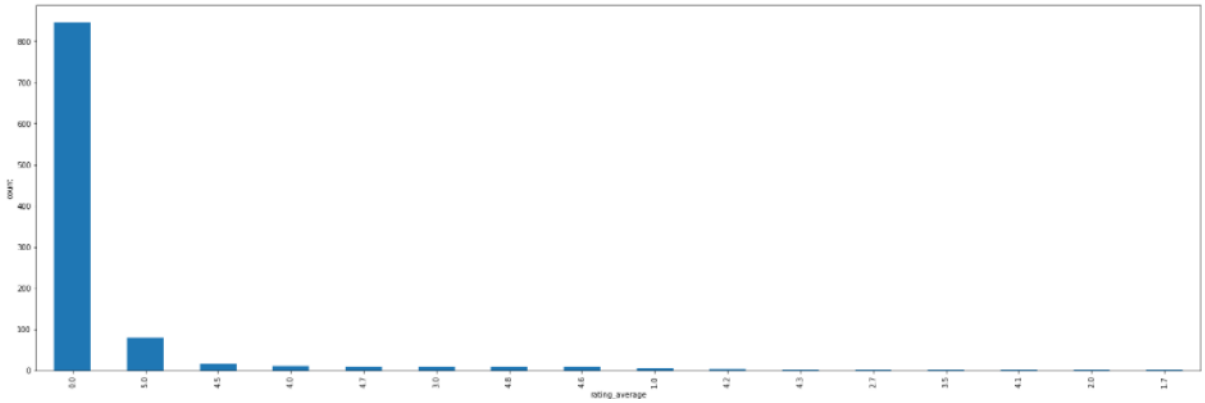
Hình 5: Biểu đồ biểu diễn số lượng đã bán và nhóm của sản phẩm



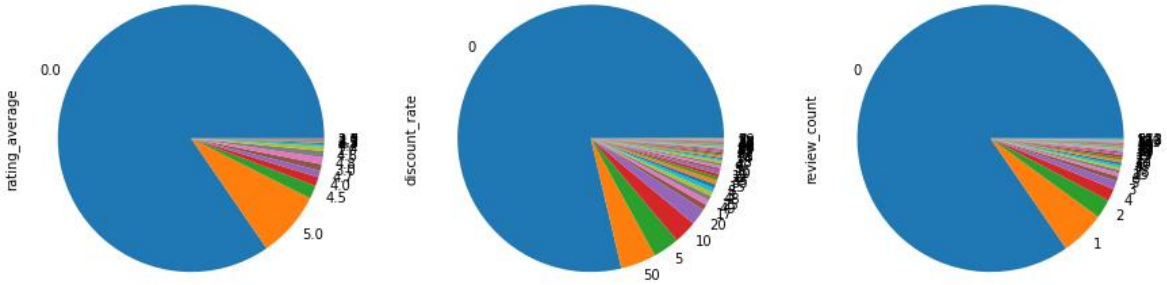
Hình 6: Biểu đồ biểu diễn điểm đánh giá và tỉ lệ discount của các sản phẩm



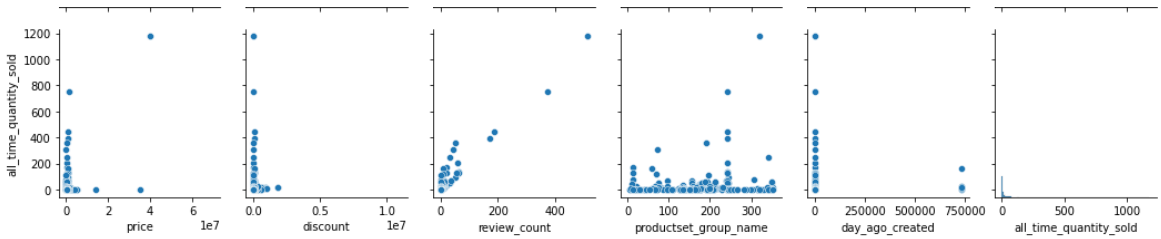
Hình 7: Biểu đồ biểu diễn số lượng đánh giá của sản phẩm



Hình 8: Biểu đồ biểu diễn số lượng điểm đánh giá của sản phẩm



Hình 9: Biểu đồ biểu diễn số lượng điểm đánh giá, discount, review



Hình 10: Các biểu đồ thể hiện độ tương quan của đặc trưng "all_time_quantity_sold" và các đặc trưng khác

Nhận xét: Qua các biểu đồ trên, có thể thấy rằng trong tổng số 1000 mẫu được xét, thì có rất ít sản phẩm được người dùng để lại bình luận, đánh giá nên số lượng các giá trị là 0 thường chiếm rất nhiều. Và cũng rất ít sản phẩm được giảm giá (discount bằng 0), chủ yếu là giữ nguyên giá gốc không thay đổi.

3. Trích xuất đặc trưng

Qua các biểu đồ thể hiện độ tương quan giữa các đặc trưng (Hình 10), em chọn ra các đặc trưng như là price, review_count, day_ago_created, discount, productset_group_name để áp dụng vào bài toán dự đoán số sản phẩm đã bán được (all_time_quantity_sold)

Từ bảng mô tả dữ liệu (Hình 4), ta thấy rằng đặc trưng "day_ago_created" có 1 dữ liệu trống nên sẽ điền giá trị median vào. Đặc trưng "all_time_quantity_sold" có đến 755 giá trị trống nên sẽ thay thế bằng giá trị 0 vào (0 sản phẩm đã được bán)

Sau khi đã điền các dữ liệu trống, em tiến hành chuẩn hóa dữ liệu áp dụng MinMaxScaler và RobustScaler

	price	discount	review_count	productset_group_name	day_ago_created	all_time_quantity_sold
0	1537000	658000	9	243	185.0	15.0
1	1207000	0	372	243	238.0	752.0
2	39900000	0	513	319	734.0	1175.0
3	1086000	121000	38	243	301.0	71.0
4	699000	0	1	243	965.0	10.0

Hình 11: Dữ liệu ban đầu

	price	discount	review_count	productset_group_name	day_ago_created	all_time_quantity_sold
0	0.021933	0.059818	0.017544	0.680672	0.000251	0.012766
1	0.017211	0.000000	0.725146	0.680672	0.000323	0.640000
2	0.570791	0.000000	1.000000	0.893557	0.000995	1.000000
3	0.015480	0.011000	0.074074	0.680672	0.000408	0.060426
4	0.009943	0.000000	0.001949	0.680672	0.001308	0.008511

Hình 12: Dữ liệu sau khi chuẩn hóa MinMax

	price	discount	review_count	productset_group_name	day_ago_created	all_time_quantity_sold
0	4.146875	658000.0	9.0	0.206096	-0.139031	15.0
1	3.115625	0.0	372.0	0.206096	-0.018234	752.0
2	124.031250	0.0	513.0	0.647315	1.112251	1175.0
3	2.737500	121000.0	38.0	0.206096	0.125356	71.0
4	1.528125	0.0	1.0	0.206096	1.638746	10.0

Hình 13: Dữ liệu sau khi chuẩn hóa Robust

4. Mô hình hóa dữ liệu

Với bài toán này, em chọn ra 2 mô hình để xây dựng gồm LinearRegression và RandomForestRegressor.

4.1. LinearRegression

LinearRegression (Hồi quy tuyến tính) là một phương pháp thống kê để hồi quy dữ liệu với biến phụ thuộc có giá trị liên tục trong khi các biến độc lập có thể có một trong hai giá trị liên tục hoặc là giá trị phân loại. Nói cách khác hồi quy tuyến tính là một phương pháp để dự đoán biến phụ thuộc (Y) dựa trên giá trị của biến độc lập (X).

Bộ tham số gồm:

Biến phụ thuộc: Y

Biến độc lập: X

4.2. RandomForestRegressor

RandomForestRegressor là một tập hợp mô hình, rất có hiệu quả cho các bài toán phân loại vì nó huy động cùng lúc hàng trăm mô hình nhỏ hơn bên trong với quy luật khác nhau để đưa ra quyết định cuối cùng. Mỗi mô hình con có thế mạnh điểm yếu khác nhau nên ta sẽ có cơ hội phân loại chính xác hơn so với khi sử dụng bất kì một mô hình đơn lẻ nào.

Như tên gọi, RandomForest dựa trên cơ sở:

Random: Tính ngẫu nhiên

Forest: nhiều cây quyết định (decision tree)

Bộ tham số gồm:

Biến phụ thuộc: Y

Biến độc lập: X

4.3. Đánh giá mô hình

Trước tiên là chọn ra các biến phụ thuộc và biến độc lập để áp dụng vào cả 2 mô hình.

Ở đây các biến độc lập X là: review_count, price, day_ago_created, gr_name, discount

Biến phụ thuộc Y là: all_time_quantity_sold

Sau đó chia các biến đó thành một tập huấn luyện và kiểm thử với tỉ lệ 7:3

Sau khi tách sẽ tiến hành đào tạo mô hình trên tập huấn luyện và thực hiện các dự đoán trên tập kiểm thử

Trong bài toán này em dựa vào R2 Score để đánh giá độ chính xác của 2 mô hình trên.

Công thức tính: $R^2 = \frac{TSS-RSS}{TSS}$

Trong đó:

TSS: là một phép đo tổng biến thiên trong tỉ lệ đáp ứng / biến phụ thuộc (Y) và có thể được coi là số lượng biến thiên vốn có trong đáp ứng trước khi hồi quy được thực hiện.

RSS: đo lường lượng biến đổi còn lại không giải thích được sau khi thực hiện hồi quy

TSS – RSS: đo lường mức độ thay đổi trong đáp ứng được giải thích (hoặc loại bỏ) bằng cách thực hiện hồi quy

R^2 : giao động từ 0 đến 1, chỉ ra mức độ mà biến phụ thuộc có thể dự đoán được

Áp dụng vào 2 mô hình và 2 cách chuẩn hóa dữ liệu, ta thu được kết quả sau:

	NaN	MinMax	Robust
Linear	0.849246	0.849246	0.849246
RandomForest	0.918484	0.915997	0.913749

Hình 14: Bảng đánh giá độ chính xác của các mô hình

Từ bảng trên ta thấy được rằng với mô hình RandomForest sẽ cho ra kết quả R^2 cao hơn mô hình Linear nhưng không thể xác định được với cách chuẩn hóa dữ liệu nào thì mô hình sẽ chính xác hơn bởi vì RandomForest là mô hình dựa trên cơ sở là tính ngẫu nhiên nên sẽ cho ra những kết quả khác nhau ở mỗi lần chạy chương trình. Nhưng nhìn chung

với cách chuẩn hóa nào hay không thì RandomForest đều cho ra kết quả lớn hơn 91%, khá cao và có thể nói dự đoán hầu như hoàn toàn chính xác nên có thể tin cậy được.

5. Kết luận

Để có thể giải quyết bài toán lần này, em đã tự mình đi thu thập dữ liệu thô từ trang tiki.vn và tiến hành xử lý, làm sạch dữ liệu để dễ dàng thực hiện bước mô tả dữ liệu một cách trực quan và dễ tiếp cận hơn.

Tiếp đến là dựa vào mô tả của các đặc trưng mà chọn ra những đặc trưng phù hợp để đưa vào làm biến độc lập và biến phụ thuộc và tiến hành xây dựng 2 mô hình để dự đoán độ tin cậy của bài toán.

Và kết quả là bài toán đã đạt được độ tin cậy khá cao, với R^2 score lên đến hơn 91%

Để bài toán tiếp tục phát triển và hoạn thiện thì tốt nhất vẫn là thu thập thêm thật nhiệu dữ liệu hơn nữa và xây dựng thêm nhiều mô hình khác để thực hiện đánh giá, chứ không chỉ phụ thuộc vào 2 mô hình hồi quy cơ bản này.

6. Tài liệu tham khảo

SV liệt kê các TLTK đã trích dẫn (cite) trong báo cáo tại đây.

[1]: https://viblo.asia/p/linear-regression-hoi-quy-tuyen-tinh-trong-machine-learning-4P856akRlY3, ngày truy cập 28/6/2022

[2]: https://viblo.asia/p/phan-lop-bang-random-forests-trong-python-djeZ1D2QKWz, ngày truy cập 28/06/2022

[3]: https://chidokun.github.io/2020/05/crawl-tiki-products/, ngày truy cập 20/06/2022