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

KHÁM PHÁ DỮ LIỆU BÁN HÀNG CỦA DOANH NGHIỆP CHUYÊN NHẬN ORDER

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Nguyễn Diệu Phương	21520091
2	Nguyễn Thị Huyền Trang	21520488
3	Nguyễn Thị Mai Trinh	21522718

TP. HỒ CHÍ MINH – 12/2023

1.GIỚI THIỆU

Xã hội ngày càng phát triển, chất lượng cuộc sống ngày càng nâng cao, cùng với đó hình thức mua sắm qua các dịch vụ mua hàng hộ từ nước ngoài và vận chuyển về Việt Nam đã trở thành lựa chọn phổ biến của người tiêu dùng. Tuy nhiên, không chỉ dừng lại ở việc cung cấp dịch vụ mua hàng, doanh nghiệp càng quan tâm đến làm thế nào để hiểu rõ hơn về thị trường, dự đoán tiềm năng của sản phẩm và đáp ứng đúng mong đợi của khách hàng. Điều này đã thúc đẩy chúng tôi lựa chọn thực hiện đề tài **Khám phá bộ dữ liệu bán hàng của doanh nghiệp chuyên nhận order**.

Cụ thể, nhóm thu thập bộ dữ liệu đơn đặt hàng của một doanh nghiệp với 13 thuộc tính liên quan sản phẩm khách muốn đặt hàng. Trong đề tài, nhóm đã xây dựng và áp dụng các phương pháp phân tích, thăm dò, xây dựng mô hình trên bộ dữ liệu đã thu thập để dự báo doanh thu và tiềm năng của sản phẩm, nhằm đáp ứng nhu cầu ngày càng cao trong việc hiểu rõ về thị trường mua sắm và đáp ứng mong muốn của người tiêu dùng.

Bộ dữ liệu được thu thập một cách minh bạch từ nguồn dữ liệu thô [1] (dạng excel) của shop Hàng Mỹ Nga Hoàng, một doanh nghiệp nhỏ chuyên nhận order hàng hiệu giá rẻ. Đây là một bộ dữ liệu do nhóm tự thu thập, chứa thông tin về sản phẩm của các đơn đặt hàng từ khách hàng. Để đảm bảo tính minh bạch và chất lượng của dữ liệu, chúng tôi đã nhận được sự đồng ý từ chủ doanh nghiệp trước khi thu thập. Bộ dữ liệu này được phân tích và thiết kế mà không có sự tham khảo hay sử dụng bất kỳ báo cáo, đề tài nghiên cứu nào khác.

Dựa trên một số mô hình học máy cơ bản như Linear Regression, Ridge Regression, Lasso Regression, Random Forest, Decision Tree, Gradient Boosting Regressor, KNeighborsRegressor và tiến hành thực hiện thực nghiệm với những bộ tham số khác nhau để tìm ra bộ tham số tối ưu nhất ứng với từng mô hình bằng *GridSearchCV* và tiến hành kiểm chứng chéo với 5 folds để xây dựng mô hình dự đoán doanh thu theo ngày chính xác nhất. Với kết quả thu được R^2vượt 0.9, nhóm tin rằng có thể phần nào hỗ trợ doanh nghiệp xác định rõ mong muốn của nhóm khách hàng tiềm năng, dự đoán doanh thu sắp tới, từ đó đưa ra hướng đi và chiến lược phù hợp.

2. MÔ TẢ BỘ DỮ LIỆU

- **Tên:** Bộ dữ liệu đơn đặt hàng của shop Hàng Mỹ Nga Hoàng [1].
- **Ngày thu thập:** 23/09/2023
- **Phương thức:** Bộ dữ liệu được nhóm thu thập trực tiếp từ nguồn dữ liệu đơn đặt hàng của một doanh nghiệp nhỏ chuyên nhận order các món hàng hiệu giá rẻ.
- **Mô tả:** Bộ dữ liệu ban đầu gồm 1966 điểm dữ liệu với 13 thuộc tính. Loại bỏ các thuộc tính không có ý nghĩa với bài toán thì bộ dữ liệu còn 9 biến: 5 biến phân loại, 4 biến kiểu số. Bộ dữ liệu phản ánh đặc điểm sản phẩm mà doanh nghiệp giao dịch trong khoảng thời gian từ ngày 14/01/2023 đến 22/09/2023.

Thuộc tính	Kiểu dữ liệu	Mô tả

date	datetime	Ngày khách hàng yêu cầu đơn đặt hàng
name_web	object	Tên website cửa hàng mua hàng
link_product	object	Liên kết sản phẩm
subject	object	Đối tượng sử dụng: W (woman: nữ), M (Man: nam), U (Unisex: phi giới tính)
color	object	Màu của sản phẩm
category	object	Danh mục của sản phẩm
amount	int	Số lượng sản phẩm order
purchase_price_unit	float	Giá mua của một sản phẩm
sale_price_unit	float	Giá bán của một sản phẩm

Bảng 1. Bảng Mô Tả Các Biến Trong Bộ Dữ Liệu Thô

3.PHƯƠNG PHÁP PHÂN TÍCH

Hình 1. Quy Trình PTDL

1.1. Xác định bài toán

"Order hàng hộ" đã dần phát triển thành một ngành nghề với thu nhập đáng kể. Không chỉ dừng lại ở việc cung cấp dịch vụ mua hàng, doanh nghiệp càng quan tâm đến làm thế nào để hiểu rõ hơn về thị trường, dự đoán tiềm năng của sản phẩm và đáp ứng đúng mong đợi của khách hàng.

Để giải quyết những thách thức này, nhóm chúng tôi đã đặt ra bài toán 'Khám phá bộ dữ liệu bán hàng của doanh nghiệp chuyên nhận order' với bộ dữ liệu về loại hình kinh doanh này.

1.2. Thu thập dữ liệu

Nhóm đã tiến hành thu thập bộ dữ liệu từ shop Hàng Mỹ Nga Hoàng một cách cẩn thận và minh bạch. Trong quá trình này, chúng tôi không áp dụng các công cụ kỹ thuật phức tạp mà thay vào đó, nhóm sử dụng công cụ Excel để lấy dữ liệu thô.

Sau khi thu thập, chúng tôi đã tiến hành một số xử lý sơ bộ dữ liệu để đảm bảo tính chính xác và sẵn sàng cho quá trình xử lý tiếp theo và phân tích. Các bước xử lý sơ bộ bao gồm việc format lại các cột để đảm bảo sự thống nhất, sắp xếp đúng theo dòng để tạo ra một cơ sở dữ liệu có tổ chức, và loại bỏ các ký tự rác,... giúp làm sạch dữ liệu và loại bỏ nhiễu từ thông tin thu thập.

Quá trình xử lý sơ bộ này không chỉ giúp chúng tôi tạo ra một bộ dữ liệu có cấu trúc mà còn đảm bảo tính đồng nhất của thông tin. Điều này là quan trọng để chúng tôi có thể thực hiện phân tích và xây dựng mô hình dự đoán với độ tin cậy cao, hỗ trợ quyết định kinh doanh và đáp ứng mục tiêu nghiên cứu của chúng tôi một cách hiệu quả.

1.3. Tiền xử lý dữ liệu

1.3.1. Chuẩn hóa

Dữ liệu khi sau thu thập còn tồn tại nhiều giá trị ngoại lai, dữ liệu bất hợp lý, hoặc thiếu sót thuộc tính để phục vụ cho bài toán. Vì thế, sau khi thăm dò dữ liệu, nhóm đã tiến hành chuẩn hóa như sau *(*có sự tư vấn ý kiến từ phía doanh nghiệp)*

- **Thuộc tính 'date':** Tạo các thuộc tính 'year', 'month', 'day' từ 'date', chúng đều chỉ thời gian đặt hàng, các thuộc tính 'year', 'month', 'day' phục vụ quá trình gom nhóm và tính toán.
- **Thuộc tính 'subject':** Thống nhất đưa về 3 phân loại là W (woman: nữ), M (Man: nam), U (Unisex: cả hai).
- **Thuộc tính 'color':** Có khá nhiều giá trị không phù hợp, ngoài ra có những màu là nhánh phụ của màu khác, hoặc nhiều tên gọi khác nhau của cùng 1 màu. Vì vậy, sau khi đưa chúng về format chuẩn mà nhóm quy định (tiếng anh viết thường), nhóm tiến hành xuất ra 1 cột màu mới:

  'color1': chuyển toàn bộ màu về nhánh chính của nó.
- **Thuộc tính 'category':** Xuất hiện khá nhiều lỗi nhập liệu, và phân loại sản phẩm chưa đủ khái quát. Sau khi tiến hành quan sát, nhóm tiến hành chuyển cột category thành 3 cột ('category', 'category1' và 'category2') với quy tắc tương tự cột color, cụ thể:

  'category': Bao gồm các category có dạng category1_category

  'category1': Chuyển các category về nhóm phù hợp: 'bag', 'dress', 'shoes', …

  'category2': Phân loại tổng quát hơn, với các phân loại: 'bag', 'clothing', 'shoes', 'accessory', 'cosmetic', 'other', nan

- **Thuộc tính 'amount'**: Dữ liệu có khá nhiều điểm không hợp lệ (0, nan). Sau khi quan sát, nhóm nhận ra có thể xử lý theo 2 trường hợp:
    - Các dữ liệu có 'purchase_price' là null hoặc '0' là những trường hợp đã bị hủy hoặc chưa đạt được thỏa thuận, hoặc hết hàng: Nhóm quyết định loại bỏ các điểm dữ liệu này vì nó không phù hợp với bài toán.
    - Các trường hợp 'amount' = 0 hoặc rỗng sẽ tiến hành điền 'amount' = 1.

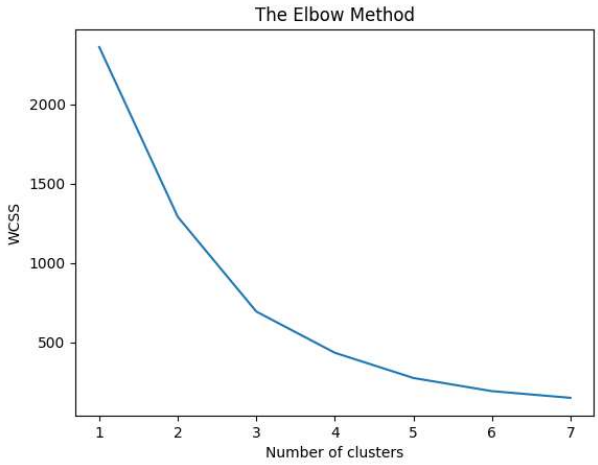
Ngoài ra, để phù hợp đề tài, nhóm bổ sung thêm các thuộc tính:

- **'sale_price'** : Tổng giá bán được. CT = 'sale_price_unit' * 'amount'
- **'purchase_price'**: Tổng giá mua hàng. CT = 'purchase_price_unit' * 'amount'
- **'revenue'** : Tổng lợi nhuận. CT = 'sale_price' - 'purchase_price'
- **'%revenue_unit'** : Phần trăm lãi trên một sản phẩm

CT = ('sale_price_unit' - 'purchase_price_unit') / 'sale_price_unit' * 100

1.3.2. Điền khuyết

Bên cạnh đó, dữ liệu còn nhiều giá trị mang giá trị null cần tiến hành điền khuyết. Do đây là dữ liệu thực tế nên nhóm tiến hành tìm ra các quy luật cũng như các mối liên quan đến các biến khác. Một số biến khuyết không nhiều và có thể dễ dàng nhận ra mối quan hệ bằng mắt nhìn như 'subject', 'link_product'. Bên cạnh đó, biến color và color1 khuyết khá nhiều và khó có thể phán đoán giá trị cần điền, chính vì vậy nhóm tiến hành phương pháp KNN (K-Nearest Neighbors) để điền khuyết các dữ liệu còn thiếu và xác định cụm (n = 4) bằng phương pháp The Elbow Method như Hình 2

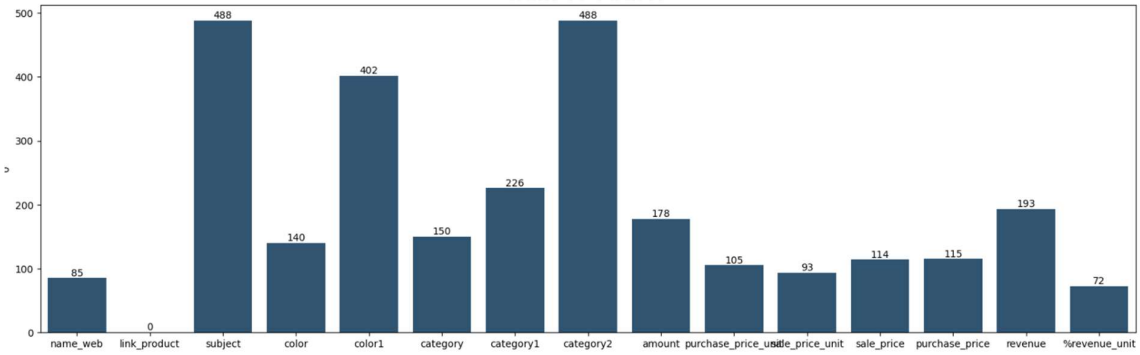


Hình 2. Phương pháp The Elbow xác định số cụm

Sau khi loại bỏ hết các cột không có giá trị, điền khuyết bổ sung dữ liệu, bộ dữ liệu hiện tại có tổng cộng 1640 điểm dữ liệu với 19 thuộc tính, bổ sung thêm 10 thuộc tính (‘year’, ‘month’, ‘da’, ‘color1’, ‘category1’, ‘category2’, ‘sale_price’, ‘purchase_price’, ‘revenue’, ‘%revenue_unit’) so với dữ liệu trong Bảng 1.

1.4. Phân tích thăm dò

Sau khi xử lí và làm sạch, bộ dữ liệu bao gồm 9 biến phân loại và 10 biến kiểu số, với 1640 điểm dữ liệu về đơn order hàng không chứa giá trị bị khuyết. Tất cả biến đều phân bố không đều, chứa nhiều giá trị ngoại lại và không ảnh hưởng lớn đến giá trị mục tiêu. Các biến có mức ảnh hưởng nhất là “***name_web***” với p_value bằng ***0.0124*** và biến “***sale_price***” có độ tương tương quan đạt ***0.7293***. Các biến còn lại đều có mực ảnh hưởng rất thấp hoặc không ảnh hưởng đến biến mục tiêu “***revenue***”.



Hình 4. Số lượng biến outlier của từng thuộc tính

Nguyên nhân của việc các biến không ảnh hưởng đến giá trị mục tiêu và chứa nhiều giá trị ngoại lệ là do bộ dữ liệu chỉ chứa thông tin sản phẩm khách hàng order. Nghĩa là, khách hàng tìm kiếm và order sản phẩm ngẫu nhiên chứ không tập trung vào một thị trường nhất định. Tuy nhiên, có một vài lựa chọn nhận được sự ưu ái từ các khách hàng tiềm năng của doanh nghiệp. Nên nhóm tiến hành xem xét, đánh giá và chọn lọc các điểm dữ liệu phù hợp, nhằm nâng cao mức ảnh hưởng đến biến mục tiêu và tăng xác suất dự đoán cho mô hình.

Để đánh giá các biến dữ liệu với biến mục tiêu, ta tiến hành chia bộ dữ liệu theo kiểu dữ liệu gồm dữ liệu phân loại và dữ liệu kiểu số và chỉ tập trung vào các biến tiềm năng (name_web, color, subject, category2, amount, sale_price, purchase_price)

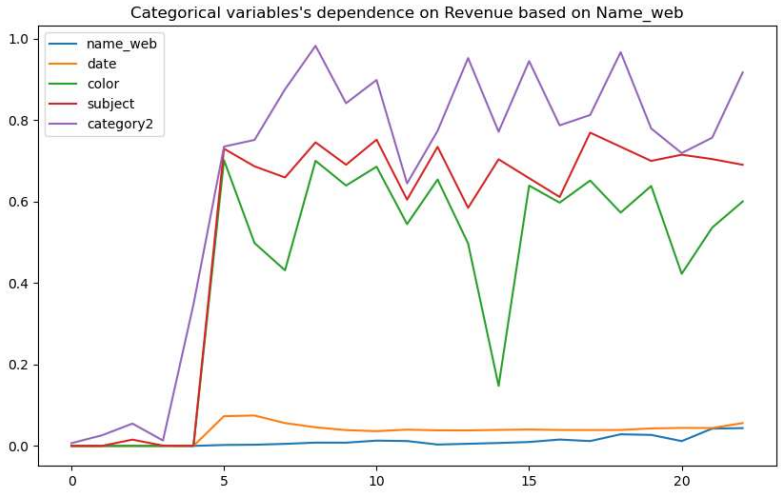
1.4.1. Biến kiểu phân loại

Với các biến kiểu phân loại, thông qua phân tích phương sai (ANOVA) và thu được kết quả như Bảng 2, ta nhận thấy các biến này không có sự khác biệt quá nhiều hay ta nói các biến này không ảnh hưởng nhiều đến biến mục tiêu. Trong đó, biến link_product như một biến phân biệt cho sản phẩm, date là biến thời gian và nhóm các biến (color, color1), (category, category1, category2) là các biến phụ thuộc lẫn nhau. Chính vì vậy nhóm tiến hành tập trung thăm dò và nâng cao mức ảnh hưởng của các biến name_web, color, subject, category2.

name_web	date	color	subject	category2	category1	link_product	category	color1
0.0122	0.0373	0.4606	0.6761	0.8831	0.9949	0.9973	0.9999	0.9999

Bảng 2. Đánh giá tương quan của các biến phân loại đến biến mục tiêu

Nhóm lần lượt tiến hành thực nghiệm so sánh mức ảnh hưởng của các biến phân loại lên giá trị mục tiêu thông qua phân tích phương sai (ANOVA) khi lần lượt loại bỏ các giá trị ngoại lai nhất của biến như Hình 5, và tiến hành giữ các điểm giá trị hữu ích sao cho tổng mức ảnh hưởng và mức ảnh hưởng riêng lẻ cao nhất mà không làm dữ liệu thay đổi hay mất mát quá nhiều. Kết quả thu được sau khi so sánh Bảng 3, Bảng 6 chứng minh hướng nhóm đi là đúng đắn.



Hình 5. ĐGTQ của biến phân loại đến biến mục tiêu khi thay đổi Name_web

name_web	**date**	**color**	**subject**	**category2**	**name**	**total**
1.224e-02	0.0373	0.4606	0.6761	0.8831	dataset	2.0693
1.423e-08	0.0010	0.0023	0.0180	0.9674	dataset11	0.9888

Bảng 3. So sánh tương quan của các biến phân loại đến biến mục tiêu trước và sau khi xử lí (trong đó dataset là bộ dữ liệu gốc, dataset11 là bộ dữ liệu sau khi xử lí)

1.4.2. Biến kiểu số

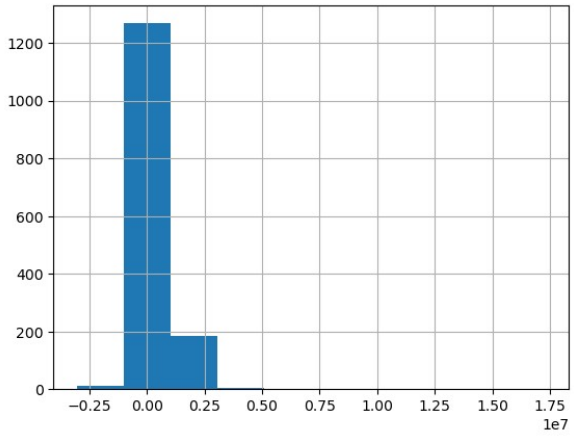
Từ Bảng 3 và quá trình EDA ta nhận thấy: tương tự biến phân loại các biến kiểu số đều cho hệ số tương quan trung bình và yếu, cao nhất là biến sale_price đạt **0.738** sau khi xử lí các biến ngoại lai phân loại. Tuy nhiên, để nâng cao độ chính xác cho mô hình, nhóm tiếp tục tiến hành EDA và thực nghiệm nhằm tìm ra và giữ lại những điểm dữ liệu có giá trị, không gây nhầm lẫn cho mô hình.

name	**dataset**	**coef**	**P_value**
amount	dataset	0.4388	3.8812e-78

	dataset11	0.4881	7.5706e-89
purchase_price	dataset	0.4368	2.1564e-77
	dataset11	0.5064	1.3794e-96
sale_price	dataset	0.7293	2.6347e-272
	dataset11	0.7380	4.5562e-253
sale_price_unit	dataset	0.3689	4.9556e-54
	dataset11	0.3684	1.8246e-48

Bảng 6. So sánh ĐTQ của biến kiểu số đến BMT khi xử lý biến phân loại (dataset11)

Bên cạnh đó, 'revenue' tuy là biến mục tiêu nhưng chứa rất nhiều giá trị ngoại lai và phân bố không đều do 2 nguyên nhân chính là (1) lợi nhuận âm (revenue <0) do khách hủy đơn (sale_price = 0) hoặc mua cao hơn báo giá (sale_price>0) (2) lợi nhuận rất lớn khi nhận được đơn đặt hàng lớn và mua trúng dịp giảm giá mạnh. Cả hai trường hợp đều không thường xuyên nên nhóm tiến hành thực nghiệm liệu có nên thu gọn ngưỡng của 'revenue' hay không. Kết quả thu được cho thấy độ tương quan giữa các biến cải thiện đáng kể khi lợi nhuận nằm trong khoảng (0, 5e6).



Hình 6. Biểu đồ histogram của biến revenue

Sau khi tiến hành xử lý, kết quả thu được vượt ngoài kỳ vọng ban đầu. Mặc dù xử lí ưu tiên cho biến kiểu số, nhưng theo kết quả Bảng 9 mức ảnh hưởng của biến phân đã được nâng lên. Đối với biến kiểu số, hệ số tương quan đều tăng cao ít nhất 0.08 (biến amount tăng 0.37), các biến kiểu số đều đạt tương quan trung bình tốt (trên 0.7).

Tuy nhiên, trong quá trình xử lý nhằm đảm bảo tính tương quan tổng thể để tăng hiệu suất cho mô hình, biến 'sale-price-unit' từ tương quan yếu thành không tương quan với biến revenue. Nhưng xét về ý nghĩa, biến 'sale_price_unit' là đơn giá của 1 sản phẩm trong khi mục tiêu bài toán là dự đoán lợi nhuận hằng ngày (bao gồm rất nhiều đơn hàng với nhiều đơn giá khác nhau). Chính vì thế, nhóm cho rằng đây là dấu hiệu cho thấy quá trình xử lý đã diễn ra đúng hướng.

name	dataset	coef	P_value
amount	dataset	0.4388	3.8812e-78
	dataset11	0.4881	7.5706e-89
	dataset18	0.8073	1.7812e-319
purchase_price	dataset	0.4368	2.1564e-77
	dataset11	0.5064	1.3794e-96
	dataset18	0.7038	6.3810e-208
sale_price	dataset	0.7293	2.6347e-272
	dataset11	0.7380	4.5562e-253
	dataset18	0.8165	0.0000e+00
sale_price_unit	dataset	0.3689	4.9556e-54
	dataset11	0.3684	1.8246e-48

Bảng 8. So sánh ĐTQ của biến kiểu số đến BMT sau khi xử lý biến số (dataset18)

Nam_web	date	color	subject	Category2	name	Total
1.2239e-02	0.0373	0.4606	0.6762	0.8831	dataset	2.0693
1.4229e-08	0.0010	0.0023	0.0180	0.9674	dataset11	0.9888
7.8720e-38	0.0000	0.0000	0.0110	0.6985	dataset18	0.7096

Bảng 9. So sánh ĐTQ của biến phân loại đến BMT sau khi xử lý biến số (dataset18)

1.5. Phát triển Mô hình

1.5.1. Chọn ra các biến ảnh hưởng đến biến mục tiêu:

Sau khi phân tích thăm dò thu được kết quả như Bảng 8, Bảng 9, nhóm đã chọn ra được các biến có khả năng ảnh hưởng đến biến mục tiêu '**revenue**' như sau:

1. **name_web:** Tên trang web đặt mua sản phẩm (adidas, amazon, ashford, macys, saksoff5th, ssense)
2. **color:** Màu sắc của sản phẩm (25 màu sắc khác nhau)
3. **subject:** Đối tượng sử dụng (M, W, U)
4. **category2:** Danh mục sản phẩm (shoes, clothing, bag, accessories, cosmetic)
5. **amount:** Số lượng sản phẩm order
6. **sale_price:** Tổng giá bán ra của đơn hàng
7. **purchase_price:** Tổng giá mua vào của đơn hàng

1.5.2. Bộ dữ liệu doanh thu theo ngày

Để phù hợp với mục đích bài toán và các mô hình học máy hiện nay, nhóm tiến hành tổng dữ liệu đơn hàng theo ngày, được bộ dữ liệu 'day_dataset' với 47 cột và 138 điểm dữ liệu (ngày). Đối với biến phân loại, nhóm tính hành tính số lượng sản phẩm liên quan trong ngày và dummies thành các cột có dạng amount_

name	coef	p_value
amount	0.9550	1.0504e-73
sale_price	0.9472	4.4214e-69
purchase_price	0.9089	1.6090e-53
amount_w	0.8551	1.2454e-40
amount_black	0.7250	8.7913e-24
amount_shoes	0.6979	1.8545e-21
amount_macys	0.6203	4.91994e-16

Bảng 10. Top 7 biến ảnh hưởng nhất đến lợi nhuận mỗi ngày

Các biến trong bộ dữ liệu ngày càng có độ tương quan cao với lợi nhuận trong ngày "revenue". Quan sát top 7 biến có độ tương quan cao nhất với lợi nhuận ở Bảng 9, ta dễ dàng nhận ra nhóm khách hàng tiềm năng của doanh nghiệp này tập trung order các sản phẩm của nữ giới, màu đen, giày, đến từ trang web macys.

1.5.3. Thử nghiệm trên các mô hình dự đoán khác nhau

Sau khi đã chọn ra được tập biến có độ tương quan cao với biến mục tiêu, nhóm sẽ thực hiện thử nghiệm trên các mô hình tuyến tính khác với những bộ tham số khác nhau để tìm ra bộ tham số tối ưu nhất ứng với từng mô hình bằng GridSearchCV. Để đảm bảo quá trình EDA và xử lý dữ liệu hiệu quả, nhóm tiến hành so sánh kết quả mô hình bằng thang đo R^2 trên ba bộ dữ liệu (1) dataset (chưa xử lý biến), (2) dataset11 (đã xử lý biến phân loại), (3) dataset18 (đã xử lý biến kiểu số và phân loại)

Mô hình	Dataset	Dataset11	Dataset18
KNeighborsRegressor	0.4830	0.7864	0.7930
GradientBoostingRegressor	0.6042	0.8692	0.8579
RandomForestRegressor	0.4913	0.7947	0.90467
DecisionTreeRegressor	0.4271	0.1078	0.6930
LinearRegression	0.4958	0.3378	0.8770
LassoRegression	0.6126	0.0547	0.8771
RidgeRegression	0.5239	0.5490	0.9093

Bảng 11. So sánh kết quả các mô hình tuyến tính trên 3 tập dữ liệu

Để có cái nhìn chính xác hơn về khả năng của các mô hình, nhóm tiến hành sử dụng tập dữ liệu đã được xử lý dataset18 cùng các siêu tham số đã tìm được bằng GridSearchCV, thực hiện kiểm chứng chéo với 5 folds. Kết quả thu được được như sau:

Mô hình	Không kiểm chứng chéo	Kiểm chứng chéo
KneighborsRegressor	0.7930	0.8217
GradientBoostingRegressor	0.8695	0.8243
RandomForestRegressor	0.9008	0.9076
DecisionTreeRegressor	0.6823	0.8071
LinearRegression	0.8770	0.8770
LassoRegression	0.8771	0.8771
RidgeRegression	0.9093	0.9093

Bảng 12. Kết quả các mô hình khi sử dụng kiểm chứng và không kiểm chứng chéo

1.5.4. Nhận xét

- Mô hình RandomForest và Ridge thể hiện hiệu suất vượt trội (hơn 0.9 điểm R^2)
    - RidgeRegression là một kỹ thuật hồi quy tuyến tính mở rộng, trong đó các trọng số của các biến độc lập được thu nhỏ và giữ lại các đặc trưng để giảm hiện tượng đa cộng tuyến. Điều này giúp cải thiện hiệu suất của mô hình, làm giảm sai số dự báo và tăng độ chính xác. RidgeRegression cũng có thể xử lý được các bài toán có số lượng biến độc lập lớn hơn số lượng phụ thuộc tương tự bộ dữ liệu này.
    - Random Forest Regression cho hiệu suất cao vì nó kết hợp nhiều cây quyết định, được xây dựng trên các tập con ngẫu nhiên của dữ liệu và các thuộc tính. Điều này giúp giảm thiểu sai số và tránh hiện tượng quá khớp (overfitting) khi so sánh với một cây quyết định đơn lẻ. Ngoài ra, Random Forest Regression còn có thể đánh giá được mức độ quan trọng của các thuộc tính và loại bỏ những thuộc tính không có tác dụng, có khả năng trích xuất ra sự tương quan ở cả những biến không tuyến tính, xuất hiện rất nhiều trong tập biến huấn luyện.
- So sánh kết quả khi sử dụng và không sử dụng kiểm chứng chéo không khác biệt quá nhiều cho thấy:
    - Mô hình có hiệu suất khá tốt và độ phức tạp phù hợp với tập dữ liệu, không bị quá khớp (overfitting) hay dưới khớp (underfitting).

- Tập dữ liệu có kích thước lớn và phân bố đồng đều, không có nhiều nhiễu hay ngoại lệ.

4. KHÓ KHĂN & THỬ THÁCH

Quá trình phân tích và khám phá bộ dữ liệu đơn đặt hàng không phải là một công việc dễ dàng, mà có thể đối mặt với nhiều khó khăn đáng kể về nhiều mặt.

(1) Bản thân doanh nghiệp chưa hiểu rõ tầm quan trọng của dữ liệu, chưa đầu tư và còn sơ sài, thiếu sót trong quá trình lưu trữ thông tin đơn đặt hàng. Do đó, xuất hiện nhiều lỗi nhập liệu và mất đồng nhất giữa các ghi chép

Ví dụ: các từ 'black', 'đen', 'Black', 'den', ... đều là chỉ màu đen

(2) Bản chất bộ dữ liệu chứa các thông tin chủ quan từ phía khách hàng, dẫn đến việc mất tính bao quát và cân bằng của dữ liệu. Chính vì vậy, hầu hết các biến đều bị lệch và chứa nhiều giá trị ngoại lai làm giảm độ tương quan giữa các biến.

(3) Nhóm chưa hiểu biết sâu về kiến thức kinh tế, dẫn đến việc khó khăn trong quá trình xác định mong muốn của doanh nghiệp, khách hàng cũng như các yếu tố có ảnh hưởng đến mục tiêu nằm ngoài bộ dữ liệu.

5.KẾT LUẬN

Ngày nay, các dịch vụ mua hàng hộ nước ngoài đã trở thành lựa chọn phổ biến của người tiêu dùng. Chính vì vậy càng có nhiều doanh nghiệp lựa chọn đi theo hướng này. Song, không chỉ dừng lại ở việc cung cấp dịch vụ mua hàng, doanh nghiệp càng quan tâm đến làm thế nào để hiểu rõ hơn về thị trường, dự đoán tiềm năng của sản phẩm và đáp ứng đúng mong đợi của khách hàng. Điều này thúc đẩy nhóm tiến hành khám phá, phân tích và tìm kiếm những insight phù hợp cho quá trình dự đoán doanh thu, hỗ trợ doanh nghiệp đưa ra phán đoán và hướng phát triển cho tương lai

Cụ thể, quá trình này bao gồm các bước thu thập, tiền xử lý, phân tích thăm dò, xây dựng và tìm kiếm mô hình tối ưu nhất cho bộ dữ liệu. Kết quả dự đoán trên các mô hình tuyến tính đều cho kết quả R^2 đạt trên 0.69. Trong đó, mô hình RandomForest và Ridge thể hiện hiệu suất vượt trội đạt trên 0.9 điểm R^2. Các mô hình cho ra hiệu suất khá tốt và có độ phức tạp phù hợp với tập dữ liệu. Tập dữ liệu sau khi được xử lý được cho là có kích thước lớn, phân bố đồng đều và không có nhiều nhiễu hay ngoại lệ.

Tóm lại, nhóm xây dựng thành công đề tài "Khám phá dữ liệu bán hàng của doanh nghiệp chuyên nhận order". Với kết quả thu được vượt 90%, nhóm tin rằng có thể phần nào hỗ trợ doanh nghiệp xác định rõ mong muốn của nhóm khách hàng tiềm năng, dự đoán doanh thu sắp tới, từ đó đưa ra hướng đi và chiến lược phù hợp. Trong tương lai, nhóm sẽ nhìn nhận sâu sắc hơn, phát triển mô hình sâu hơn để cải thiện khả năng dự đoán và ứng dụng thực tiễn trong doanh nghiệp.

TÀI LIỆU THAM KHẢO

[1] Link bộ dữ liệu thô:
https://docs.google.com/spreadsheets/d/1IS97VPig1YYRGw4ZYbYftBrBmfWJN6PSSKNeqyJOWNE/edit?usp=sharing (23/09/2023)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Diệu Phương	- Phân tích thăm dò + Xây dựng mô hình
2	Nguyễn Thị Mai Trinh	- Tiền xử lí + Phân tích thăm dò
3	Nguyễn Thị Huyền Trang	- Thu thập dữ liệu + Tiền xử lí + Slide
4	Chung	- Viết báo cáo