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

DỰ ĐOÁN TIỀN BOA MỖI CHUYẾN ĐI CỦA TAXI MÀU XANH LÁ TẠI THÀNH PHỐ NEW YORK

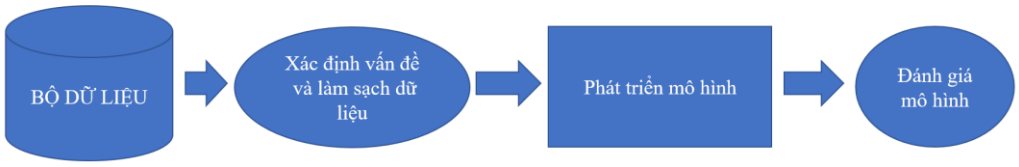
Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Lâm Gia Huy	18520832
2	Nguyễn Xuân Vinh	18521655

TP. HỒ CHÍ MINH – 12/2020

1. GIỚI THIỆU

- Trong bài báo cáo này, chúng tôi sẽ đi phân tích để làm sạch bộ dữ liệu "2020 Green Taxi Trip Data (January - June)" và phát triển phương pháp phù hợp cho bài toán dự đoán tiền boa trên bộ dữ liệu này.
- Để thực hiện, chúng tôi đã tìm hiểu sâu về ý nghĩa của các đặc trưng có trong bộ dữ liệu từ đó đưa ra những ràng buộc để loại bỏ hoặc thay thế những dữ liệu nhiễu, đồng thời chúng tôi có phát triển thêm các đặc trưng mới từ các đặc trưng đã có . Sau đó, chúng tôi sử dụng kĩ thuật phân tích thăm dò để kiểm tra sự tương tác giữa đặc trưng mục tiêu (tiền boa) với các đặc trưng còn lại, từ đó chọn ra những đặc trưng có tương quan với đặc trưng mục tiêu để phát triển phương pháp hồi quy cho việc dự đoán tiền boa. Bên cạnh đó, từ các đặc trưng phát triển thêm, chúng tôi vạch ra các hướng để phát triển bài toán. Cuối cùng, chúng tôi sẽ so sánh các phương pháp vơi nhau thông qua các thông số R2, K-fold validation để đưa ra phương pháp phù hợp nhất có thể cho bài toán của chúng tôi.
- Và phương pháp hồi quy đa thức với mô hình Polynomial Linear Regression cho chúng tôi kết quả khả quan nhất. Với kết quả R2 đạt đến 0.85 và K-fold validation là 0.79.

2. NỘI DUNG



Hình 1. Quy trình PTDL.

2.1 Bộ dữ liệu

- Tên: 2020 Green Taxi Trip Data (January - June) **[1]**
- Ý nghĩa: Đây là hồ sơ chuyến đi,do Nhà cung cấp dịch vụ công nghệ taxi xanh (TSP) thực hiện. Mỗi hàng đại diện cho một chuyến đi duy nhất trên taxi màu xanh lá cây vào năm 2020 (6 tháng đầu). Hồ sơ chuyến đi bao gồm các trường ghi lại ngày / giờ đón và trả khách, địa điểm đón và trả khách, khoảng cách chuyến đi, giá vé chia thành từng khoản, loại giá, loại thanh toán , và số lượng hành khách do tài xế báo cáo.
- Số lượng đặc trưng: 20
- Số lượng mẫu: 1.225.889

Số thứ tự	Đặc trưng	Ý nghĩa	Kiểu dữ liệu	Miền giá trị
1	VendorID	Mã nhà cung cấp bản ghi Taxicab Passenger Enhancement Programs (TPEP)	category	1= Creative Mobile Technologies, LLC 2= Curb Mobility (formerly VeriFone Inc)
2	tpep_pickup_datetime	Thời điểm đón khách được đồng hồ xe ghi nhận	datetime	Từ 22:06:48 ngày 31/12/2008 đến 04:48:19 ngày 14/07/2020

3	tpep_dropoff_ datetime	Thời điểm đón khách được đồng hồ xe ghi nhận	datetime	Từ 23:12:08 ngày 31/12/2008 đến 05:05:37 ngày 14/07/2020
4	store_and_fwd _flag	Cờ cho việc kiểm tra liệu dữ liệu của chuyến đi có được lưu trong bộ nhớ xe trước khi đưa đến nhà cung cấp, vì xe không có kết nối đến máy chủ	category	Y= lưu trữ và chuyển tiếp chuyến đi N= không lưu trữ và chuyển tiếp chuyến đi
5	RatecodeID	Mã giá vé (có hiệu lực vào cuối chuyến đi)	category	1= Standard rate 2= JFK (Các chuyến đi giữa Manhattan và Sân bay John F. Kennedy (JFK)) 3= Newark (Các chuyến đi đến Sân bay Newark (EWR)) 4= Nassau or Westchester(Giá ngoài Thành phố đến Nassau hoặc Westchester.) 5= Negotiated fare (Giá vé Cố định

				Thương lượng Ngoài Thành phố) 6= Group ride(Giá vé theo nhóm) 99: Nhập liệu sai
6	PU_LocationID	Mã khu vực (nơi đón khách)	category	Tương ứng với trường location_id trong NYC Taxi Zones (1 đến 265)
7	DO_LocationID	Mã khu vực (nơi trả khách)	category	Tương ứng với trường location_id trong NYC Taxi Zones (1 đến 265)
8	passenger_count	Số lượng khách trên xe (Đây là giá trị do người lái xe nhập và có thể không phản ánh chính xác số lượng hành khách trên một chuyến đi)	category	Nhóm 1 đến 9 người
9	trip_distance	Khoảng cách chuyến đi **(** tính bằng dặm) được báo cáo bằng đồng hồ xe taxi	float	Liên tục từ -33.69 đến 149267.7
10	fare_amount	Giá tiền tính theo quãng đường (dặm)	float	Liên tục từ -210 đến 753

		được ghi nhận trên đồng hồ tốc độ		
11	extra	Các phụ phí thêm: 1$ cho qua đêm(8 giờ tối đến 6 giờ sáng) Mã giá 1 : 0.5$ cho giờ cao điểm (từ 4 giờ chiều đến 8 giờ tối các ngày trong tuần, không bao gồm ngày nghỉ hợp pháp) Mã giá 2: 4,50$ giờ cao điểm (4 giờ chiều đến 8 giờ tối các ngày trong tuần, trừ ngày lễ). Mã giá 3: 17.5$ phí Netwalk và phí cầu đường Mã giá 4: phụ thu phí cầu đường đối với các chuyến đi qua các Cựu chiến binh Vịnh Cross và Cầu tưởng niệm Marine	float	Liên tục -4.5 đến 16.74

		Parkway-Gil Hodges		
12	mta_tax	Thuế di chuyển cho các phương tiện trong đô thị	category	0 : Không tính thuế 0.5: Thuế di chuyển mặc định -0.5, 3.55
13	tip_amount	Tiền mà khách boa cho tài xế, được tự động tính vào thẻ tín dụng (Không bao gồm hình thức toán tiền mặt)	float	Liên tục từ -10.56 đến 641.2
14	tolls_amount	Tổng số tiền của tất cả các khoản phí phải trả trong chuyến đi.	float	Liên tục từ -6.12 đến 96.12
15	ehail_fee	Phí đặt chỗ	category	Không có dữ liệu
16	improvement_surcharge	Phụ phí cải thiện chuyến đi	category	0: Không tính phi 0.3: phụ phí cải thiện mặc định -0.3
17	total_amount	Tổng số tiền được tính cho hành khách(Không bao gồm tiền boa)	float	Liên tục từ -210.3 đến 753.8
18	payment_type	Hình thức thanh toán	category	1= Credit card(tín dụng) 2= Cash(tiền mặt)

				3= No charge(miễn phí) 4= Dispute(thỏa thuận) 5= Unknown 6= Voided trip(chuyến đi bị hủy)
19	trip_type	Loại hình chuyến đi	category	1 , 2 (Chưa được cập nhật)
20	congestion_surcharge	Phụ phí tắc nghẽn (cho tất cả các chuyến đi bắt đầu, kết thúc hoặc đi qua Manhattan ở phía nam của Đường 96.)	category	0: Không tính phí 0.75: phụ phí chung 2.75: phụ phí cho taxi xanh lá 2.5: phụ phí cho taxi vàng - 2.75

Bảng 1: Mô tả các đặc trưng của bộ dữ liệu

2.2 Tiền xử lý dữ liệu

2.2.1 Làm sạch dữ liệu

Theo như bảng 1, cùng với những gì tôi tìm hiểu về quy định phí taxi ở New York **[2]**, bộ dữ liệu hiện tại vẫn còn nhiều sai sót, do đó chúng tôi đã tiến hành giải quyết những sai sót đó với các hướng xử lý sau:

2.2.1.1 Loại bỏ:

- Ehail_fee: đặc trưng này hoàn toàn bị khuyết giá trị
- Trùng lặp: các mẫu trùng lặp hoàn toàn giá trị, và các đặc trưng của bộ dữ liên quan đến Mã định danh (ID)
- Thời gian trước năm 2015: vào năm 2015 phụ phí cải thiện (congestion_surcharge) đã được điều chỉnh

- Thời điểm đón khách trễ hơn thời điểm trả khách(tpep_pickup_datetime > tpep_dropoff_datetime): đồng hồ xe sẽ được làm mới sau mỗi chuyến đi, do đó thời điểm bắt đầu chuyến đi phải nhỏ thời điểm kết thúc chuyến đi.
- Cùng thời điểm đón và trả khách(tpep_pickup_datetime = tpep_dropoff_datetime) nhưng có khoảng cách(trip_distance != 0): đồng hồ xe ghi nhận và kết thúc tại một thời điểm, theo quy tắc chiếc xe sẽ không di chuyển (có thể tài xế quên làm mới đồng hồ, nên dữ liệu khoảng cách còn lưu trữ)
- Phụ phí tắc nghẽncó giá trị 2.5(congestion_surcharge = 2.5): giá trị này áp dụng cho taxi màu Vàng trong khi chúng tôi đang đề cập đến taxi màu Xanh lá.
- Mã giá có giá trị 99(RatecodeID = 99): trong quy định về phí taxi, chỉ tồn tại 6 loại giá vé từ 1 đến 6
- Thuế di chuyển trong đô thị có giá trị 3.55(mta_tax = 3.55): trong quy định về phí, thì loại thuế này chỉ có mức quy định chung là 0.5
- Tổng số tiền phải trả có giá trị nhỏ hơn 2.5 và lớn hơn bằng 0 (0 <= total_amount < 2.5): trong quy định về phí, phí khởi đầu cho mỗi chuyến đi là 2.5.

Lưu ý: Ở đây các trường hợp này do số lượng rất ít cũng như chúng tôi chưa có cơ sở chính xác để thay thế nên quyết định loại bỏ

2.2.1.2 Thay thế:

- Các đặc trưng có giá trị âm:
    - trip_distance , fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge (Dựa vào bảng 1) : các khoản phí được đề ra để tính thêm vào nên chỉ có trường hợp có tính hoặc không tính, chứ không trừ bớt ra. Và khoảng cách là một đại lượng vô hướng, chỉ có độ lớn và không thể âm.
    - Hướng giải quyết: thay thế bằng giá trị tuyệt đối vì trong quá trình nhập có thể sai sót trong việc nhập dư dấu '-'.
- Có boa(tip_amount > 0) với hình thức thanh toán tiền mặt (payment_type = 2):
    - Theo như bảng 1, tiền boa sẽ được tính trực tiếp vào thẻ tín dụng không bao gồm tiền mặt

- Hướng giải quyết: sẽ thay thế bằng Hình thức thanh toán là thẻ tín dụng (payment_type = 1) vì các mã giá vé của các mẫu này đều có Mã giá tiêu chuẩn(RatecodeID = 1) và mã này đa số đều có hình thức thanh toán là thể tín dụng

- Khác thời điểm đón và trả(tpep_pickup_datetime != tpep_dropoff_datetime) nhưng xe lại không di chuyển (trip_distance == 0):
  - Điều nay có thể xảy ra khi chuyến đi đó bị hủy bởi khách hàng, nhưng thông thường thì xe sẽ di chuyển tức có khoảng cách. Và ở đây chúng tôi sẽ quy về trường hợp thông thường nhất, tức là các dữ liệu này bị nhập sai vì số lượng mẫu khá nhiều (3.5% bộ dữ liệu)
  - Hướng giải quyết: với các trường hợp xe có di chuyển (trip_distance != 0) và có cùng địa điểm đón(PU_LocationID) và rước(DO_LocationID) với trường hợp xe không di chuyển(trip_distance), chúng tôi sẽ tính khoảng thời gian hành trình(tpep_pickup_datetime == tpep_dropoff_datetime) của cả hai trường hợp. Với thời gian hành trình của trường hợp xe có di chuyển gần nhất so với thời gian hành trình của trường hợp xe không di chuyển, chúng tôi sẽ thay thế khoảng cách chuyến đi (trip_distance) của trường hợp xe không di chuyển bằng khoảng cách chuyến đi (trip_distance) của trường hợp xe di chuyển

- Giá trị bị khuyết:
  - Nhìn chung chỉ có khoảng 15.8% mẫu bị khuyết với các đặc trưng RatecodeID, passenger_count, trip_type, congestion_surcharge, VendorID, store_and_flag_fwd, payment_type (24.95% giá trị khuyết). Con số này không quá ít, cũng không quá nhiều nên chúng tôi quyết định thay thế
  - Hướng giải quyết: đa số các đặc trưng này có kiểu dữ liệu category do đó chúng tôi sẽ điền thiếu bằng các giá trị có tần suất xuất hiện nhiều, xét theo từng Quận và tiền boa (Với quy định về phí taxi ở New York, ở một số khu vực nhất định sẽ có quy định riêng)

Lưu ý: Thuộc tính Quận là thuộc tính mà chúng tôi phát triển thêm, sẽ được đề cập đến ở phần Phát triển thêm đặc trưng của mục Tiền xử lý dữ liệu

Ràng buộc / Đặc trưng	Tip_amount = 0		Tip_amount > 0	
	PU_Borough	DO_Borough	PU_Borough	DO_Borough
VendorID	2	2	2	2
store_and_fwd_flag	N	N	N	N
passenger_count	1	1	1	1
trip_type	1	1	1 , EWR(2)	1
congestion_surcharge	0	0	0	0 , Manhattan(2.75)
RatecodeID	1	1, EWR(3)	1, EWR(5)	1, EWR(3)
payment_type	2,Staten Island, Unknown (1)	2, Staten Island (1)	1	1

Bảng 2: Giá trị có tần suất xuất hiện nhiều nhất ở mỗi đặc trưng với mỗi điều kiện tương ứng

2.2.2 Phát triển thêm đặc trưng

Mục tiêu của bài toán của chúng tôi chính là tiền boa, và tiền boa sẽ bị ảnh hưởng rất nhiều bởi yếu tố con người. Biết được điều đó, từ các đặc trưng sẵn có chúng tôi phát triển thêm những đặc trưng có liên quan đến yếu tố con người

Đặc trưng	Cách tính	Ý nghĩa	Lý do
trip_duration	tpep_dropoff_datetime - tpep_pickup_datetime	Khoảng thời gian hành trình	Dùng để phát triển đặc trưng speed

		(giây)	
speed	trip_distance / (trip_duration / 3600)	Tốc độ của xe (dặm / giờ) (Không quá 50 dặm / giờ: quy định về tốc độ tối đa tại New York)	Đa số người đi xe mong muốn đến nơi nhanh chóng, nếu đáp ứng được sẽ có tiền boa
hour	Lấy giờ trong chuỗi thời gian tpep_pickup_datetime	Khung giờ có chuyến đi	Ở các khung giờ cao điểm, nếu giúp họ đến nơi nhanh chóng có thể có tiền boa)
PU_Borough DO_Borough	Đối chiếu PU_LocationID và DO_LocationID với Borough trong bảng taxi+_zone_lookup	Các quận ở New York	Với các quận khác nhau, sẽ có kinh tế, sự phát triển khác nhau nên sẽ lượng tiền được boa cũng sẽ chênh lệch
covid	tpep_pickup_datetime < 3: 0 Ngược lại: 1	Tháng 3 là thời gian dịch bắt đầu diễn ra	Dịch làm cho kinh tế trì trệ, tiền boa cũng như chuyến đi có thể giảm

		ở New York [**3**]	trong thời gian này
airport	Đối chiếu PU_LocationID và DO_LocationID với service_zone trong bảng Taxi Zone Lookup Table[**4**] (airport: 1, no airport: 0)	Chuyến đi có địa điểm đón hoặc trả thuộc sân bay không	Thường người đến sân bay là đón người thân hoặc đi nước ngoài, đa số những người này khá giả, tiền boa có thể cao

Bảng 3: Thông tin của các đặc trưng phát triển thêm

2.2.3 Phân tich thăm dò

2.2.3.1 Đặc trưng mục tiêu Tiền boa (tip_amount)

- Giá trị là 0 (tức không có tiền boa) chiếm phần lớn bộ dữ liệu(63.88%), điều này hiện tại làm cho độ phân bố dữ liệu của đặc trưng mục tiêu không cân bằng, giảm đi độ tương quan vốn có giữa các đặc trưng còn lại với đặc trưng mục tiêu.
- Do đó, chúng tôi quyết định sẽ phân tích và thực nghiệm đồng thời trên hai bộ dữ liệu :
    - Bộ dữ liệu lớn (bao gồm tiền boa có giá trị bằng 0)
    - Bộ dữ liệu nhỏ (chỉ gồm tiền boa có giá trị khác 0)

2.2.3.2 Chọn lọc đặc trưng

- Ở bước này, chúng tôi sẽ tính toán độ tương quan (pearson) giữa các đặc trưng và lấy các đặc trưng có hệ số độ tương quan > 0.2 so với đặc trưng mục tiêu để đưa vào mô hình huấn luyện
- Với các đặc trưng định tính, chúng tôi đã tiến hành mã hóa các giá trị của chúng theo mức độ tăng dần của trung bình tiền boa theo nhóm các giá trị của đặc trưng đó, trước khi tính toán độ tương quan. Vì bài toán chúng tôi mong muốn dự đoán ra các chuyến đi ổn định về lượng tiền boa và có tiền boa cao, do đó chúng tôi đã mã hóa bằng giá trị trung bình, giá trị trung bình sẽ phản ánh bình quân tiền boa của nhóm đó cao hay thấp.

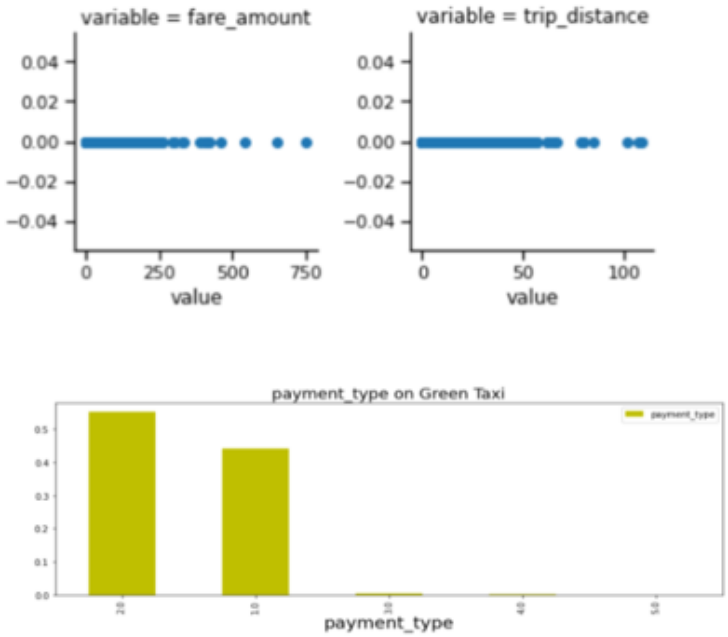
Bộ dữ liệu	total_amount	fare_amount	trip_distance	DO_LocationID	payment_type	congestion_surcharge
Lớn	**0.24**	0.07	0.08	**0.25**	**0.49**	**0.27**
Nhỏ	**0.58**	**0.38**	**0.33**	**0.29**	0.01	0.18

Bảng 4: Độ tương quan giữa đặc trưng mục tiêu (tip_amount) với các đặc trưng còn lại (chỉ xét trên các đặc trưng quan trọng tức có hệ số > 0.2) ở trên cả 2 bộ lớn và nhỏ)

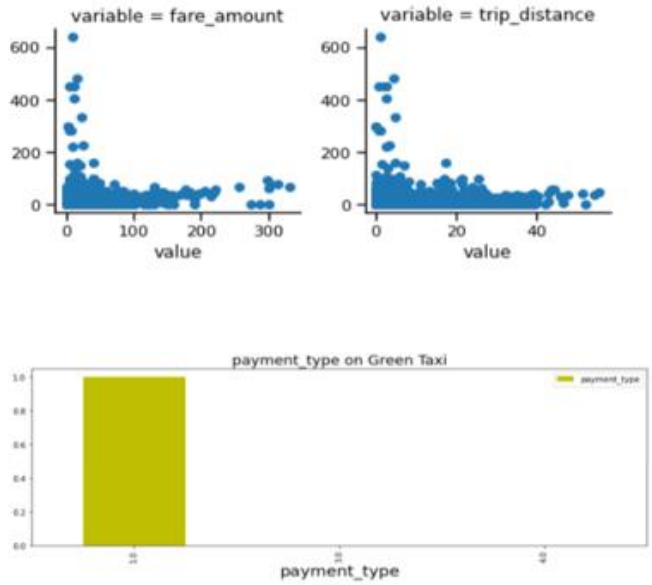
- Như vậy, các đặc trưng DO_LocationID và total_amount là hai đặc trưng luôn ảnh hưởng đến đến đặc trưng mục tiêu của chúng tôi.
- Bên cạnh có, có thêm một số đặc trưng quan trọng
    - Payment_type, congestion_surcharge đối với bộ lớn
    - Fare_amount, trip_distance đối với bộ nhỏ
- Dựa vào bảng trên, ta thấy được có sự chênh lệch khá nhiều về độ tương quan giữa 2 bộ ở các đặc trưng như fare_amount, trip_distance, payment_type.
    - Đối với các biến định lượng như fare_amount, trip_distance thì sự thay đổi này là dễ hiểu vì các dữ liệu không có tiền boa trải đều ở vùng giá trị

có tiền boa, dẫn đến mất đi sự tương quan tuyến tính giữa 2 đặc trưng này với đặc trưng mục tiêu

- Đối với đặc trưng định tính payment_type, sự thay đổi từ có tương quan (bộ lớn) thành không tương quan (bộ nhỏ) là do ở bộ lớn có sự chênh lệch tương đối giữa các hình thức với nhau , còn ở bộ nhỏ, hình thức 1 chiếm phần đại đa số với gần 100% số mẫu.



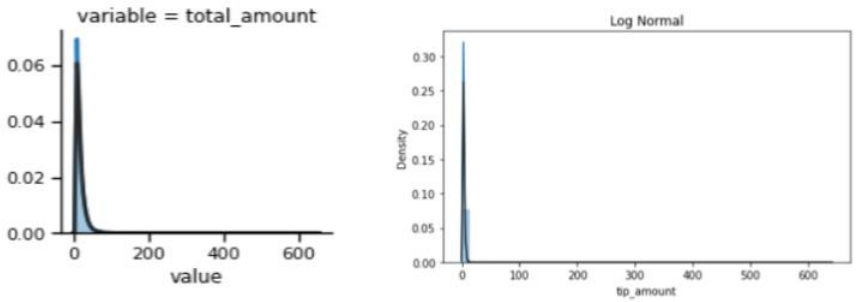
Hình 2. Tương quan giữa đặc trưng tip_amount với payment_type ở bộ lớn, trip_distance, fare_amount khi tip_amuont = 0



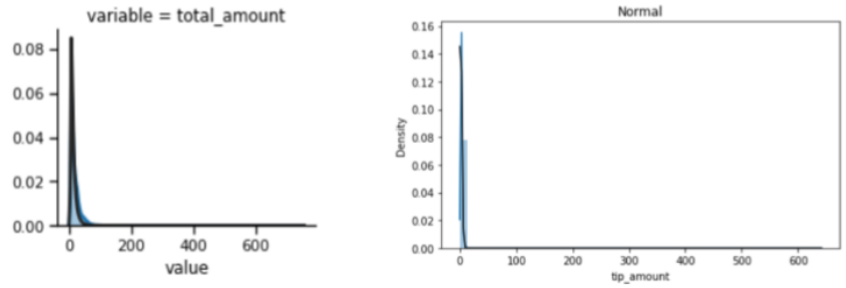
Hình 3. Tương quan giữa đặc trưng tip_amount với trip_distance fare_amount, payment_type trên bộ nhỏ

2.2.4 Chuẩn hóa dữ liệu

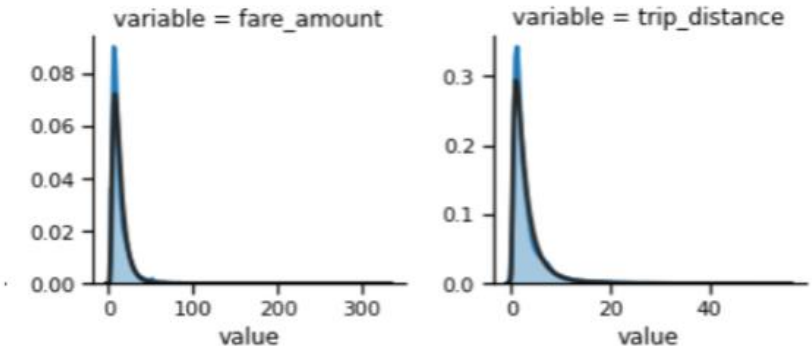
- Vì độ phân bố giá trị của các biến định lượng rộng ở miền giá trị lớn, và không đồng đều do đó, chúng tôi đã thực hiện chuẩn hóa dữ liệu về dạnh chính quy hóa có kỳ vọng bằng 0 và phương sai bằng 1 để thu nhỏ miền giá trị của đặc trưng, giúp tăng hiệu suất học cho các mô hình(dùng StandardScaler trong python)
- Với các đặc trưng trip_distance, fare_amount, total_amount chúng tôi sẽ dùng log trước khi chuẩn hóa. Đối với đặc trưng tip_amount chúng tôi sẽ chuẩn hóa cho bộ lớn và log giá trị trước khi chuẩn hóa cho bộ nhỏ.



Hình 4. Phân bố giá trị của đặc trưng tip_amount, total_amount ở bộ lớn.



Hình 5. Phân bố giá trị của đặc trưng tip_amount, total_amount ở bộ nhỏ

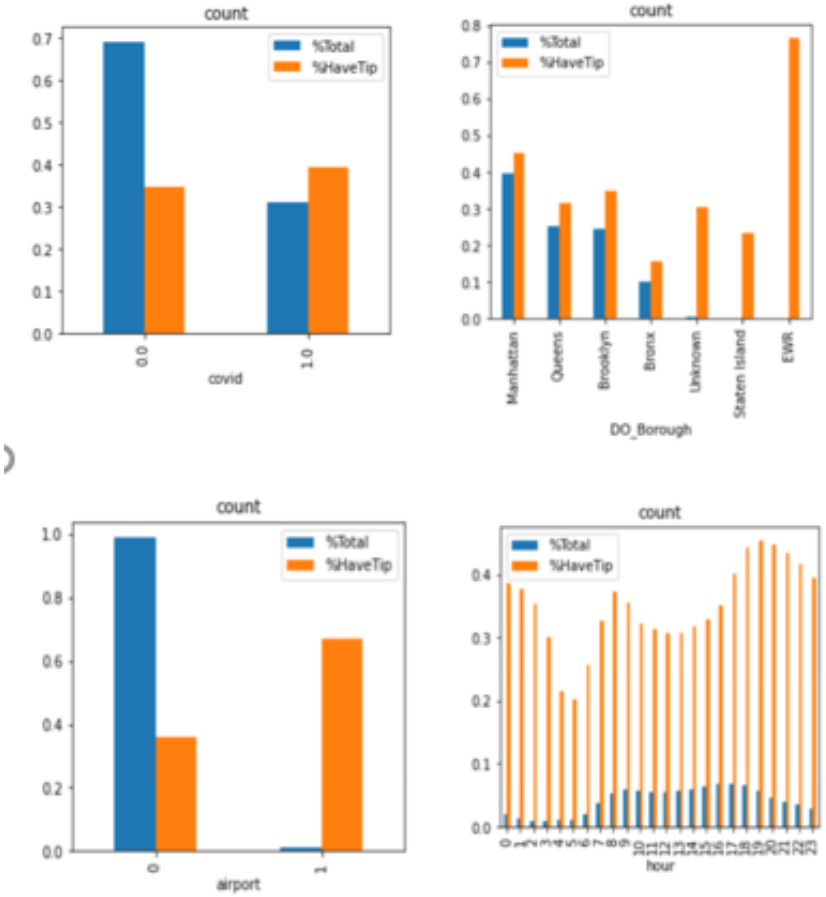


Hình 6. Phân bố giá trị của đặc trưng trip_distance, fare_amount ở bộ nhỏ

2.2.5 Các hướng khai triển bài toán

- Với các đặc trưng mà chúng tôi phát triển thêm, chúng tôi sẽ phát triển bài toán tiền boa ở cả 2 bộ dành riêng cho từng đặc trưng đó, để có thể xem xét và đánh giá cho từng nhóm giá trị ở đặc trưng đó. Cụ thể, chúng tôi sẽ xây dựng mô hình cho đặc trưng covid, airport, hour, DO_Borough.
- Ở đây, chúng tôi không phát triển cho đặc trưng PU_Borough vì đặc trưng DO_LocationID có ảnh hưởng đến đặc trưng mục tiêu cao hơn với PU_LocationID, như vậy đặc trưng DO_Borough sẽ phù hợp hơn do được phát triển từ DO_LocationID. Còn các đặc trưng được phát triển còn lại, là các đặc trưng định lượng và chúng tôi đã xét ở trên là không có tương quan với đặc trưng tiền boa (tip_amount) nên sẽ không đưa vào phát triển bài toán.

- Đặc trưng giờ(hour), ở mỗi nhóm giá trị chiếm một phần dữ liệu khá nhỏ, chưa tới 10% do đó chúng tôi sẽ gom nhóm các khung giờ với nhau như sau:
    - [0,3]: có tỉ lệ số mẫu có tiền boa lớn hơn 30%
    - [4,6]: có tỉ lệ số mẫu có tiền boa nhỏ hơn 30%
    - [7,11]: có tỉ lệ số mẫu có tiền boa dao động tăng rồi giảm nhẹ ở mức 31%.
    - [12,16]: có tỉ lệ số mẫu có tiền boa dao động quanh mức 30%
    - [17,23]: có tỉ lệ số mẫu có tiền boa cao hơn 39%



Hình 7. Sự phân bố giá trị của từng nhóm giá trị của các đặc trưng phát triển them ở bộ nhỏ so với bộ lớn

Chú thích:

- %Total: phần trăm của nhóm giá trị đó trên tổng bộ dữ liệu lớn
- %HaveTip: phần tram của nhóm giá trị đó ở bộ nhỏ so với giá trị đó ở bộ lớn

2.3 Phát triển mô hình

2.3.1 Huấn luyện mô hình

- Trong bài báo cáo này, chúng tôi đã sử dụng hai mô hình truyền thống để huấn luyện và đánh giá các trường hợp có thể xảy ra khi xử lý tiền dữ liệu bằng nhiều cách khác nhau.
- Hai mô hình mà chúng tôi sử dụng là mô hình cơ bản Linear Regression và Polynomial Linear Regression.
- Đầu vào của mô hình sẽ là giá trị của các đặc trưng: DOLocationID, trip_distance, fare_amount, total_amount, payment_type, congestion_surcharge. Đây là các đặc trưng được chúng tôi xem là quan trọng của bộ dữ liệu đối với biến taget tip_amount
- Đầu ra là: tip_amount (tiền boa của tài xế)
- Chúng tôi sẽ huấn luyện mô hình bằng cách vét cạn mọi trường hợp có thể xẩy ra để cho ra mô hình tốt nhất có thể.

2.3.2 Đánh giá mô hình

- Sau quá trình huấn luyện các mô hình, chúng tôi đã nhận được một bảng kết quả cho các mô hình tốt nhất đối với mỗi trường hợp:

Chú ý: Các mô hình được in đậm và in nghiêng là các mô hình tốt nhất so với các mô hình còn lại trong bộ dữ liệu được xét.

Bộ dữ liệu			**Model**	**Feature details**	**RMSE**	**R^2 test**	**4 fold - mean**	**5 fold - mean**	**note**
AIRPO	**0**	**FULL**	Polynomial Linear Regression	['DOLocationID', 'payment_type', 'congestion_surcharge', 'total_amount']	0.9202	0.2702	-9.6505	0.2009	test_size = 0.2, Degree = 4, num feature = 4

RT	**1**		Polynomial Linear Regression	['DOLocationID', 'payment_type', 'total_amount']	0.4715	0.7621	0.7437	0.7463	test_size = 0.2, Degree = 3, num feature = 3
	0	**TIP > 0**	Polynomial Linear Regression	['DOLocationID', 'total_amount', 'fare_amount', 'trip_distance']	0.5191	0.7311	0.6649	0.6504	test_size = 0.2, Degree = 4, num feature = 4
	1		***Polynomial Linear Regression***	***['DOLocationID', 'total_amount', 'fare_amount']***	***0.3863***	***0.8544***	***0.7926***	***0.7967***	***test_size = 0.2, Degree = 3, num feature = 3***
COVID	**0**	**FULL**	Polynomial Linear Regression	['DOLocationID', 'payment_type', 'congestion_surcharge', 'total_amount']	0.5087	0.5828	0.4224	0.4322	test_size = 0.2, Degree = 2, num feature = 4
	1		Polynomial Linear Regression	['payment_type', 'congestion_surcharge', 'total_amount']	0.9558	0.2070	-5.56E+15	-7.10E+15	test_size = 0.2, Degree = 4, num feature = 3
	0	**TIP > 0**	***Polynomial Linear Regression***	***['DOLocationID', 'trip_distance', 'fare_amount', 'total_amount']***	***0.4516***	***0.7944***	***0.6982***	***0.6983***	***test_size = 0.2, Degree = 4, num feature = 4***
	1		Polynomial Linear Regression	['trip_distance', 'fare_amount', 'total_amount']	0.5815	0.6614	0.5833	0.4669	test_size = 0.2, Degree = 4, num feature = 3

DO_BOROUGH	**Bronx**	**FULL**	Polynomial Linear Regression	['DOLocationID', 'payment_type', 'congestion_surcharge', 'total_amount']	0.7157	0.0821	-0.0369	-0.1060	test_size = 0.2, Degree = 2, num feature = 4
	Brooklyn		Polynomial Linear Regression	['DOLocationID', 'payment_type', 'congestion_surcharge', 'total_amount']	0.8090	0.0625	0.0194	-0.1559	test_size = 0.2, Degree = 2, num feature = 4
	EWR		Polynomial Linear Regression	['DOLocationID', 'payment_type', 'congestion_surcharge', 'total_amount']	0.8185	0.1725	-9.52E+24	0.3292	test_size = 0.2, Degree = 3, num feature = 4
	Manhattan		Polynomial Linear Regression	['payment_type', 'congestion_surcharge', 'total_amount']	0.5950	0.6266	-429155252	0.6030	test_size = 0.2, Degree = 3, num feature = 3
	Queens		Polynomial Linear Regression	['DOLocationID', 'payment_type', 'congestion_surcharge', 'total_amount']	0.5974	0.4754	0.3957	0.4038	test_size = 0.2, Degree = 3, num feature = 4
	Staten Island		Polynomial Linear Regression	['DOLocationID', 'payment_type', 'congestion_surcharge']	0.6284	0.3676	-2.0600	0.2261	test_size = 0.2, Degree = 3, num feature = 3

	Unknown		Polynomial Linear Regression	['DOLocationID', 'total_amount']	0.8109	0.2018	0.1056	0.1220	test_size = 0.2, Degree = 2, num feature = 2
	Bronx	**TIP > 0**	Polynomial Linear Regression	['trip_distance', 'fare_amount', 'total_amount']	0.5751	0.6719	-1.4731	-3.0489	test_size = 0.2, Degree = 4, num feature = 3
	Brooklyn		Polynomial Linear Regression	['DOLocationID', 'total_amount', 'fare_amount', 'trip_distance']	0.4593	0.7891	0.6836	0.6292	test_size = 0.2, Degree = 3, num feature = 4
	EWR		Polynomial Linear Regression	['fare_amount', 'total_amount']	0.5881	0.5636	0.6528	0.6360	test_size = 0.2, Degree = 3, num feature = 2
	Manhattan		***Polynomial Linear Regression***	***['DOLocationID', 'trip_distance', 'fare_amount', 'total_amount']***	***0.4671***	***0.7823***	***0.7607***	***0.7586***	***test_size = 0.2, Degree = 3, num feature = 4***
	Queens		Polynomial Linear Regression	['DOLocationID', 'trip_distance', 'fare_amount', 'total_amount']	0.4488	0.796	0.7267	0.7236	test_size = 0.2, Degree = 4, num feature = 4
	Staten Island		Polynomial Linear Regression	['DOLocationID', 'trip_distance', 'total_amount']	0.4586	0.754	-646.02	-692.29	test_size = 0.2, Degree = 4, num feature = 3

	Unknown		Multiple Linear Regression	['DOLocationID', 'fare_amount', 'total_amount']	0.6888	0.4984	0.3982	0.4000	test_size = 0.2, num feature = 3
HOUR	**[0,3]**	**FULL**	Polynomial Linear Regression	['DOLocationID', 'payment_type', 'congestion_surcharge', 'total_amount']	0.8956	0.2307	-59221 31292	-3.4075E+10	test_size = 0.2, Degree = 3, num feature = 4
	[4,6]		Polynomial Linear Regression	['DOLocationID', 'payment_type', 'congestion_surcharge', 'total_amount']	0.8706	0.2088	-106.99	-115.87	test_size = 0.2, Degree = 4, num feature = 4
	[7,11]		Polynomial Linear Regression	['DOLocationID', 'payment_type', 'congestion_surcharge', 'total_amount']	0.8083	0.3264	-2.0874	-6.4723	test_size = 0.2, Degree = 4, num feature = 4
	[12,16]		Polynomial Linear Regression	['payment_type', 'congestion_surcharge', 'total_amount']	0.7158	0.2312	-163041	-194431	test_size = 0.2, Degree = 4, num feature = 3
	[17,23]		Polynomial Linear Regression	['DOLocationID', 'payment_type', 'congestion_surcharge', 'total_amount']	0.6242	0.5866	0.5502	0.5477	test_size = 0.2, Degree = 4, num feature = 4

	[0,3]		***Polynomial Linear Regression***	***['DOLocationID', 'trip_distance', 'fare_amount', 'total_amount']***	***0.4651***	***0.7906***	***0.7816***	***0.7822***	***test_size = 0.2, Degree = 2, num feature = 4***
	[4,6]		Polynomial Linear Regression	['DOLocationID', 'trip_distance', 'fare_amount', 'total_amount']	0.5415	0.7026	-4.0616	0.0870	test_size = 0.2, Degree = 4, num feature = 4
	[7,11]	**TIP > 0**	Polynomial Linear Regression	['total_amount', 'fare_amount', 'trip_distance']	0.5126	0.7342	0.5960	0.6123	test_size = 0.2, Degree = 3, num feature = 3
	[12,16]		Polynomial Linear Regression	['total_amount', 'fare_amount']	0.5240	0.7293	0.6125	0.5896	test_size = 0.2, Degree = 4, num feature = 2
	[17,23]		Polynomial Linear Regression	['DOLocationID', 'trip_distance', 'fare_amount', 'total_amount']	0.4779	0.7704	0.7438	0.7439	test_size = 0.2, Degree = 4, num feature = 4

Bảng 5. Bảng kết quả mô hình trên các bộ dữ liệu

3 KẾT LUẬN

- Chúng tôi đã hoàn thành việc xử lý và phân tích bộ dữ liệu taxi màu xanh lá tại thành phố NewYork. Qua quá trình phân tích chúng tôi nhận thấy có nhiều vấn đề cần xử lý và giải quyết. Chúng tôi đã chia ra được một bộ dữ liệu nhỏ với giá trị tip_amount > 0 để tăng khả năng dự đoán cho bài toán, cùng với đó là chia bộ nhdữ liệu với các đặc trưng: Airport, covid, thời gian, khu vực. Và kết quả như mong muốn, các mô hình được huấn luyện trên bộ dữ liệu nhỏ đều cho kết quả vượt trội hơn với bộ dữ liệu ban đầu (chi tiết trên Bảng 5. Kết quả các mô hình trên bộ dữ liệu).
- Chúng tôi đã có được bốn mô hình có kết quả cao tương ứng với bốn bộ dữ liệu nhỏ(tip_amount > 0): Airport, Covid, DO_Borough và Hour.

Bộ dữ liệu	**Model**	**Feature details**	**RMSE**	**R^2 test**	**4 fold - mean**	**5 fold - mean**	**note**
Airport (1)	Polynomial Linear Regression	['DOLocationID', 'total_amount', 'fare_amount']	0.3863	0.8544	0.7926	0.7967	test_size = 0.2, Degree = 3, num feature = 3
Covid (0)	Polynomial Linear Regression	['DOLocationID', 'trip_distance', 'fare_amount', 'total_amount']	0.4516	0.7944	0.6982	0.6983	test_size = 0.2, Degree = 4, num feature = 4
DO_Borough (Manhattan)	Polynomial Linear Regression	['DOLocationID', 'trip_distance', 'fare_amount', 'total_amount']	0.4671	0.7823	0.7607	0.7586	test_size = 0.2, Degree = 3, num feature = 4
Hour ([0,3])	Polynomial Linear Regression	['DOLocationID', 'trip_distance', 'fare_amount', 'total_amount']	0.4651	0.7906	0.7816	0.7822	test_size = 0.2, Degree = 2, num feature = 4

Bảng 6. Bảng kết quả mô hình tốt nhất

TÀI LIỆU THAM KHẢO

[1] NYC Open Data, https://data.cityofnewyork.us/Transportation/2020-Green-Taxi-Trip-Data-January-June-/pkmi-4kfn (24/11/2020).

[2] NYC Taxi & Limousine Commission, https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page (9/12/2020)

[3] THE WALL STREET JOURNAL, https://www.wsj.com/articles/first-case-of-coronavirus-confirmed-in-new-york-state-11583111692 (2/12/2020)

[4] NYC Taxi & Limousine Commission, https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page (9/12/2020)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Lâm Gia Huy	• Tìm hiểu ý nghĩa các đặc trưng có trong bộ dữ liệu • Thực hiện kiểm tra và làm sạch bộ dữ liệu • Chọn các đặc trưng để phát triển bài toán tiền boa • Vạch ra nhiều kịch bản khác nhau cho bài toán • Viết báo cáo, soạn slide (Giới thiệu, Bộ dữ liệu, Tiền xử lý dữ liệu)
2	Nguyễn Xuân Vinh	• Tìm hiểu ý nghĩa các đặc trưng có trong bộ dữ liệu • Thực nghiệm các phương pháp trên bộ dữ liệu đã xử lý • Thực hiện đánh giá và so sánh các mô hình • Viết báo cáo, soạn slide(Phát triển các mô hình, Đánh giá, Kết luận)
3		