

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

□ □ □ □ □



XÂY DỰNG MÔ HÌNH DỰ ĐOÁN TIỀN TIP
DỰA TRÊN BỘ DỮ LIỆU CHICAGO TAXI
TRIPS

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Nguyễn Trọng Ân	18520434
2	Dương Văn Bình	18520505

TP. HỒ CHÍ MINH – 12/2020

1. GIỚI THIỆU

Sự ra đời của ô tô cũng tạo ra nhiều việc làm khác nhau, trong số đó có nghề lái xe taxi. Lái xe taxi đã trở thành một nghề phổ biến trong Xã hội, ngành nghề này đã xuất hiện tại nước Mỹ từ thế kỷ XX nên dữ liệu về các chuyến xe taxi phục vụ khách hàng là rất lớn và sẽ được lưu trữ lại dưới nhiều hình thức khác nhau. Dù có có được lưu trữ như thế nào thì đây cũng là nguồn dữ liệu vô cùng lớn và tiềm năng để khai thác. Trong phạm vi đồ án này chúng tôi sẽ áp dụng những kỹ thuật khai thác dữ liệu đã được học (phân tích thăm dò, trục quan dữ liệu) để tìm ra những điều thú vị ẩn dấu bên trong bộ dữ liệu về các chuyến đi taxi (Chicago Taxi Trips) và xây dựng mô hình (Hồi quy tuyến tính đơn biến, đa biến) dự đoán tiền Tip của khách hàng dành cho các tài xế taxi.

Sau quá trình tìm nghiên cứu bộ dữ liệu và tiến hành các thực nghiệm, kết quả hiện tại mà nhóm đạt được là tìm được các thuộc tính quan trọng trong bộ dữ liệu có ảnh hưởng nhiều nhất tới số tiền Tip mà tài xế nhận được cũng như xây dựng được những mô hình có kết quả khả quan trên bộ dữ liệu, đồng thời các mô hình cũng được deploy.

2. NỘI DUNG

2.1. Bộ dữ liệu

Tên bộ dữ liệu: Chicago Taxi Trips.

Nguồn: [Chicago Data Portal](#)

Số lượng điểm dữ liệu: >194 triệu dòng (trong phạm vi đồ án nhóm chỉ chọn ngẫu nhiên 200,000 điểm dữ liệu để tiến hành phân tích và thực nghiệm).

Số lượng cột: 23

Bảng 1: Codebook của bộ dữ liệu.

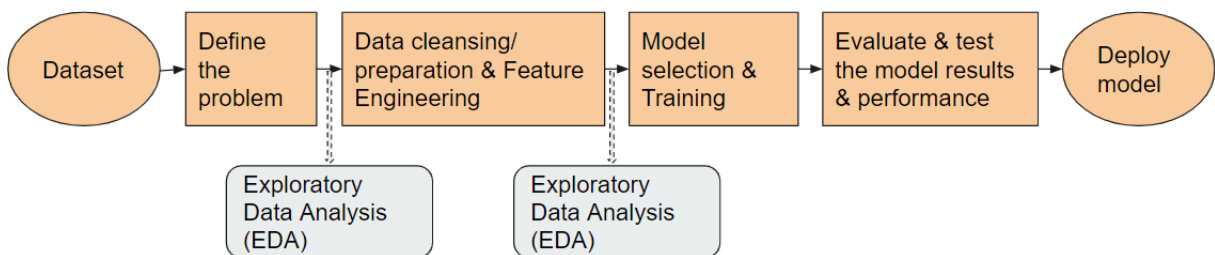
Tên cột	Miêu tả tên cột	Kiểu dữ liệu	Ý nghĩa
Trip ID	id của chuyến đi	Text	phân biệt các chuyến đi
Taxi ID	id của xe taxi	Text	phân biệt các xe taxi
Trip Start Timestamp	thời gian chuyến xe khởi hành, được làm tròn thành 15 phút gần nhất	Date & Time	thời gian và ngày tháng bắt đầu chuyến đi
Trip End Timestamp	thời gian chuyến xe kết thúc, được làm tròn thành 15 phút gần nhất	Date & Time	thời gian và ngày tháng kết thúc chuyến đi
Trip Seconds	Độ dài của chuyến đi tính bằng giây	Number	thời gian tiêu tốn

Trip Miles	Khoảng cách di chuyển tính bằng dặm	Number	khoảng cách giữa điểm đến và điểm bắt đầu
Pickup Census Tract	Census tract nơi bắt đầu chuyến đi. Để bảo mật thông tin một số chuyến đi sẽ không có thuộc tính này. Những địa điểm nằm ngoài Chicago thường không có giá trị cho thuộc tính này	Text	Census tract của khu vực nơi bắt đầu chuyến đi
Dropoff Census Tract	Census tract nơi kết thúc chuyến đi. Để bảo mật thông tin một số chuyến đi sẽ không có thuộc tính này. Những địa điểm nằm ngoài Chicago thường không có giá trị cho thuộc tính này	Text	Census tract của khu vực nơi kết thúc chuyến đi
Pickup Community Area	Khu vực dân cư nơi chuyến đi bắt đầu. Nếu địa điểm nằm ngoài Chicago thì trường này sẽ trống	Number	Số thứ tự của khu vực dân cư nơi bắt đầu chuyến đi (1-77)
Dropoff Community Area	Khu vực dân cư nơi chuyến đi kết thúc. Nếu địa điểm nằm ngoài Chicago thì trường này sẽ trống	Number	Số thứ tự của khu vực dân cư nơi kết thúc chuyến đi
Fare	Giá chuyến đi	Number	Chi phí cung cấp dịch vụ
Tips	Tiền tip khách hàng đưa cho driver. Tip bằng tiền mặt sẽ không được lưu lại	Number	Tiền Tip mà khách hàng tặng tài xế
Tolls	Phí đường bộ	Number	Thuế đường bộ
Extras	Chi phí phát sinh thêm	Number	Số tiền khách hàng phải trả thêm
Trip Total	Tổng chi phí	Number	Chi phí mà khách hàng phải trả
Payment Type	Phương thức thanh toán	Text	Phương thức thanh toán của khách hàng
Company	Hãng taxi	Text	Công ty cung cấp dịch vụ
Pickup Centroid Latitude	Vĩ độ của trung tâm nơi đón khách census tract hoặc community area nếu census tract bị ẩn. Các địa điểm nằm ngoài Chicago thường sẽ không có trường này	Number	Vĩ độ nơi bắt đầu chuyến đi
Pickup Centroid Longitude	Kinh độ của trung tâm nơi đón khách census tract hoặc community area nếu census tract bị ẩn. Các địa điểm nằm ngoài Chicago thường sẽ không có trường này	Number	Kinh độ nơi bắt đầu chuyến đi
Pickup Centroid	Địa điểm của trung tâm nơi đón khách census tract hoặc community area nếu census tract bị ẩn. Các địa điểm	Point	Toạ độ địa lý của nơi bắt đầu chuyến đi

Location	nằm ngoài Chicago thường sẽ không có trường này		
Dropoff Centroid Latitude	Vĩ độ của trung tâm nơi đến của khách census tract hoặc community area nếu census tract bị ẩn. Các địa điểm nằm ngoài Chicago thường sẽ không có trường này	Number	Vĩ độ nơi kết thúc chuyến đi
Dropoff Centroid Longitude	Kinh độ của trung tâm nơi đến của khách census tract hoặc community area nếu census tract bị ẩn. Các địa điểm nằm ngoài Chicago thường sẽ không có trường này	Number	Kinh độ nơi kết thúc chuyến đi
Dropoff Centroid Location	Địa điểm của trung tâm nơi đến của khách census tract hoặc community area nếu census tract bị ẩn. Các địa điểm nằm ngoài Chicago thường sẽ không có trường này	Point	Toạ độ địa lý của nơi kết thúc chuyến đi

Bộ dữ liệu được tổng hợp từ năm 2013 - Hiện tại, các điểm dữ liệu mới vẫn được cập nhật liên tục từng tháng.

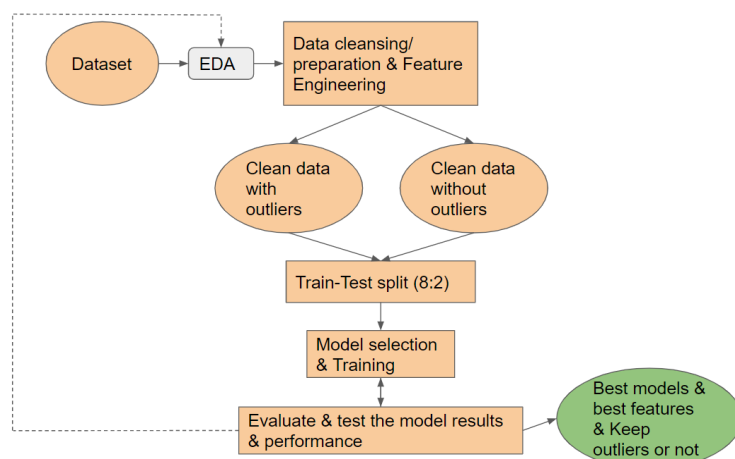
Quy trình phân tích dữ liệu:



2.2. Xác định bài toán

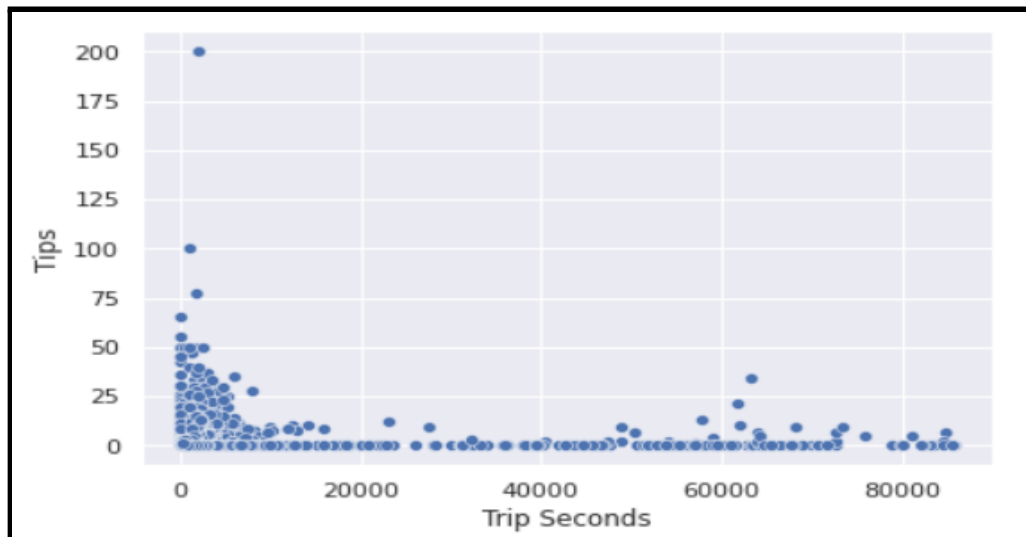
Bài toán đặt ra là việc dự đoán số tiền Tip mà một tài xế taxi có thể nhận được sau mỗi chuyến đi thông tin các chuyến đi được lưu trữ.

Quy trình thực hiện:

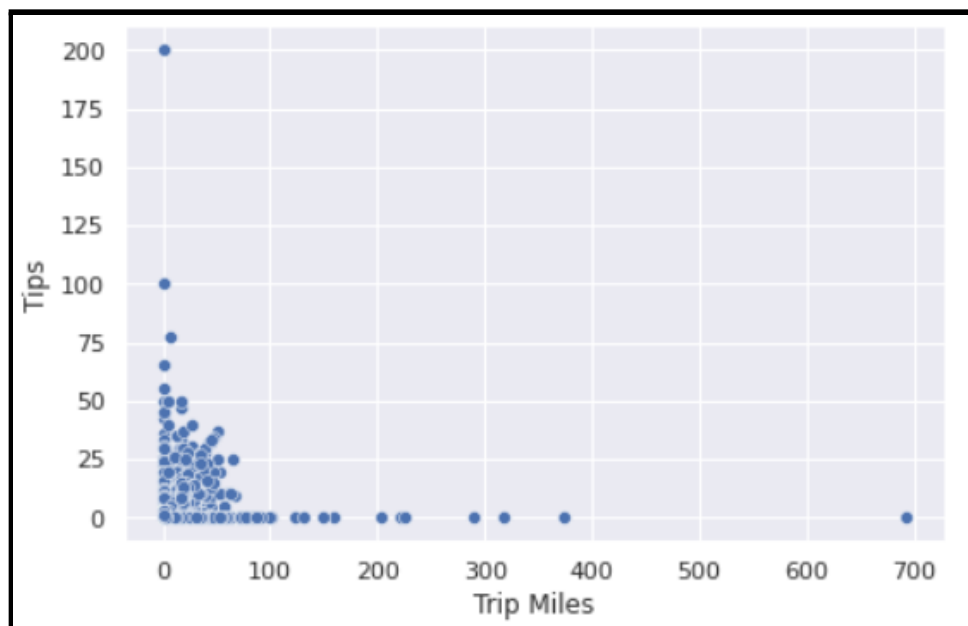


2.3. Phân tích thăm dò

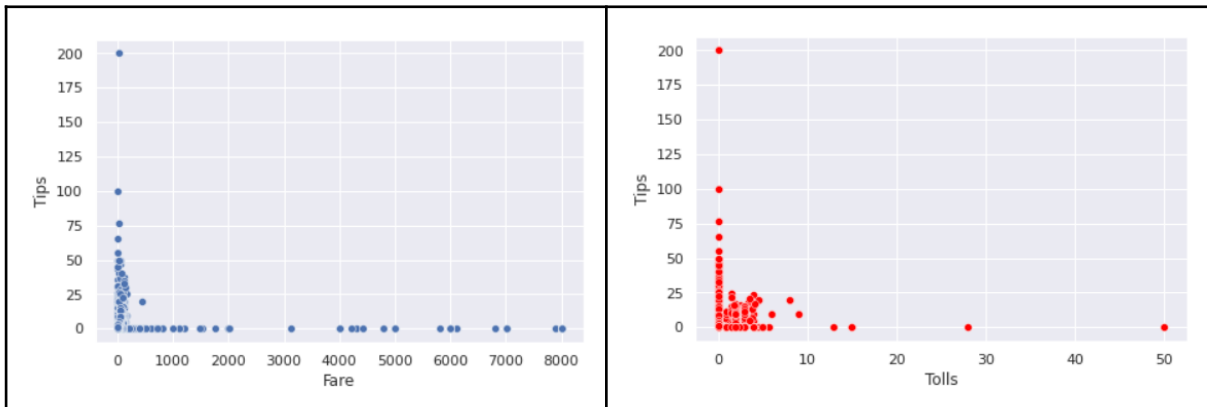
'Trip Seconds' có các điểm dữ liệu phân bố ở các miền giá trị là khá đồng đều, tuy nhiên tập chung nhiều nhất vẫn là ở khu vực nhỏ hơn 20,000 giây. Từ biểu đồ ta có thể nhận định rằng không phải chuyến đi kéo dài càng lâu thì nhận được nhiều tiền Tip. Những chuyến đi nhận được nhiều tiền Tip chủ yếu vẫn là những chuyến đi có thời gian ngắn. Cũng có một số điểm outlier cần được xử lý ở bước tiếp theo.



'Trip Miles' có khoảng giá trị phân bố chính là từ 0 tới 100, đây cũng là miền giá trị duy nhất nhận được tiền Tip. Giống như 'Trip Seconds' cũng có những giá trị outlier xuất hiện nên cũng cần phải được xử lý để không làm ảnh hưởng tới việc phát triển mô hình.

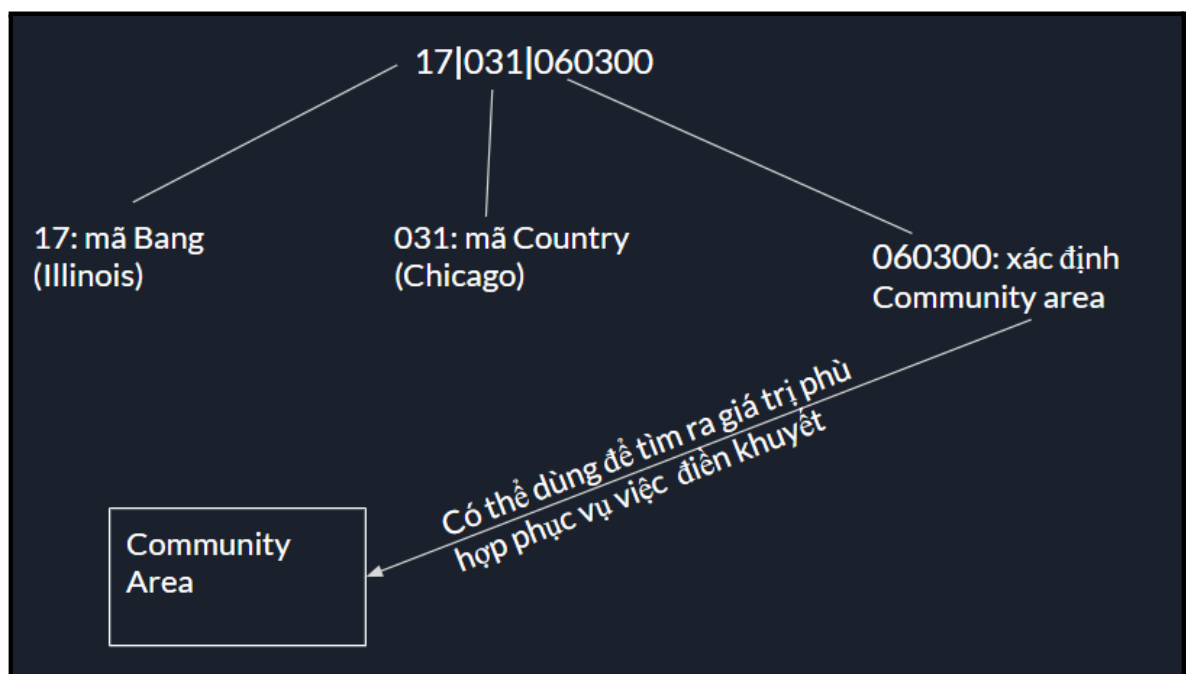


Các thuộc tính 'Fare', 'Tolls' và 'Extras' phản ánh đúng thực tế về khả năng nhận tiền Tip của tài xế trong thực tế, vì nếu những giá trị cho các thuộc tính này cao thì sẽ ảnh hưởng không nhỏ đến tổng chi phí mà khách hàng phải trả cho chuyến đi vì vậy sẽ ít khả năng nhận được tiền Tip hơn.

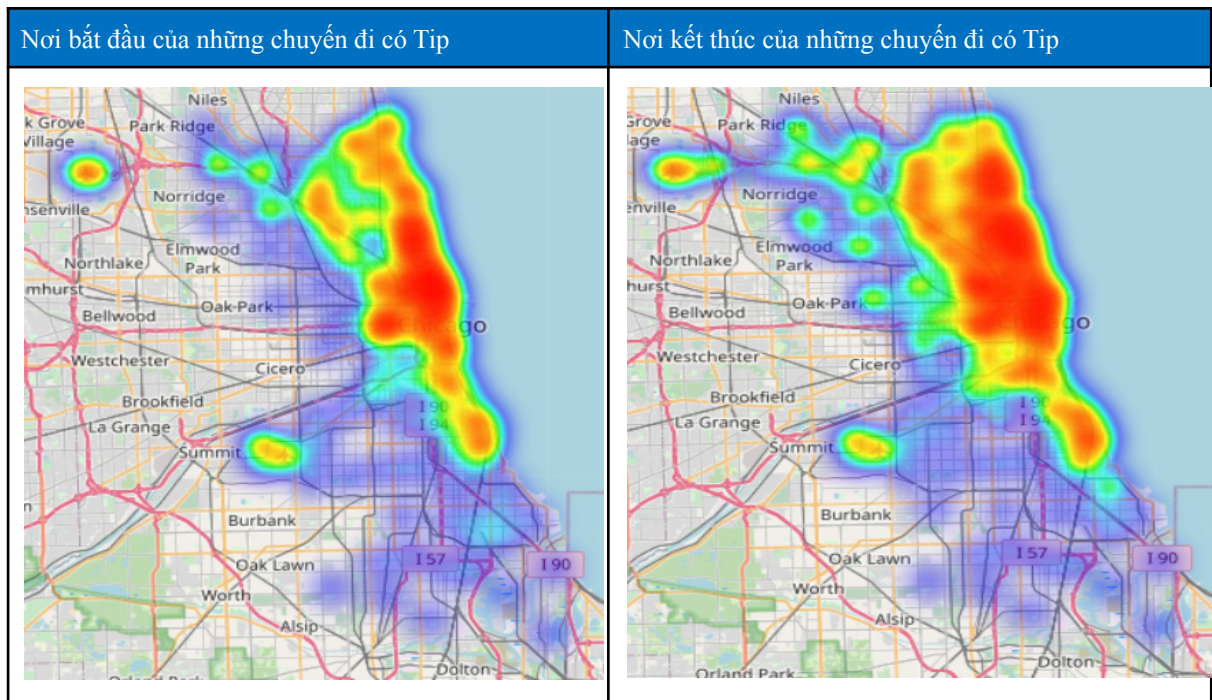


'Taxi ID' cho ta biết được rằng trong bộ dữ liệu có 3711 chiếc taxi hoạt động trong bộ dữ liệu. Và từ 'Company' ta biết được có 47 hãng taxi cung cấp dịch vụ.

Mối liên hệ giữa 'Pickup Census Tract' và 'Dropoff Census Tract' với các Community Area tương ứng.

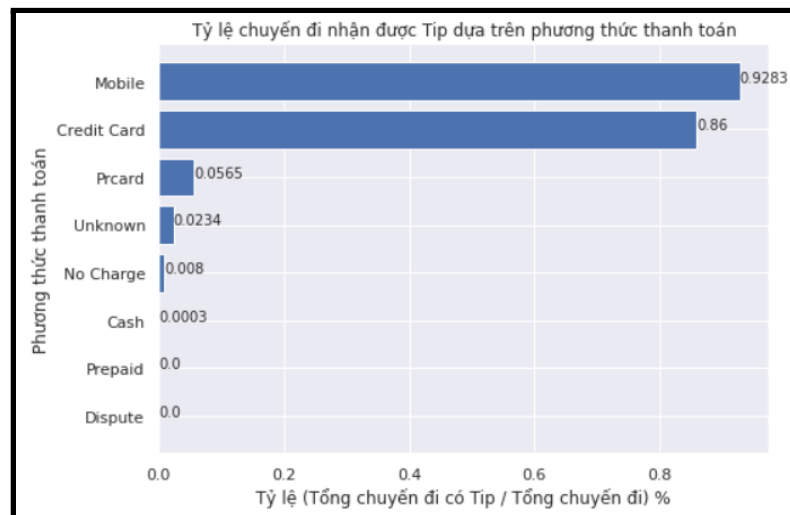


Từ tọa độ ta có biểu đồ heatmap cho thấy khu vực có nhiều chuyến đi và khu vực đến cho thấy khả năng nhận được tiền Tip ở các khu vực địa lý là khác nhau hay nói cách khác là các khu vực Community Area có ảnh hưởng tới việc nhận được Tip của tài xế.

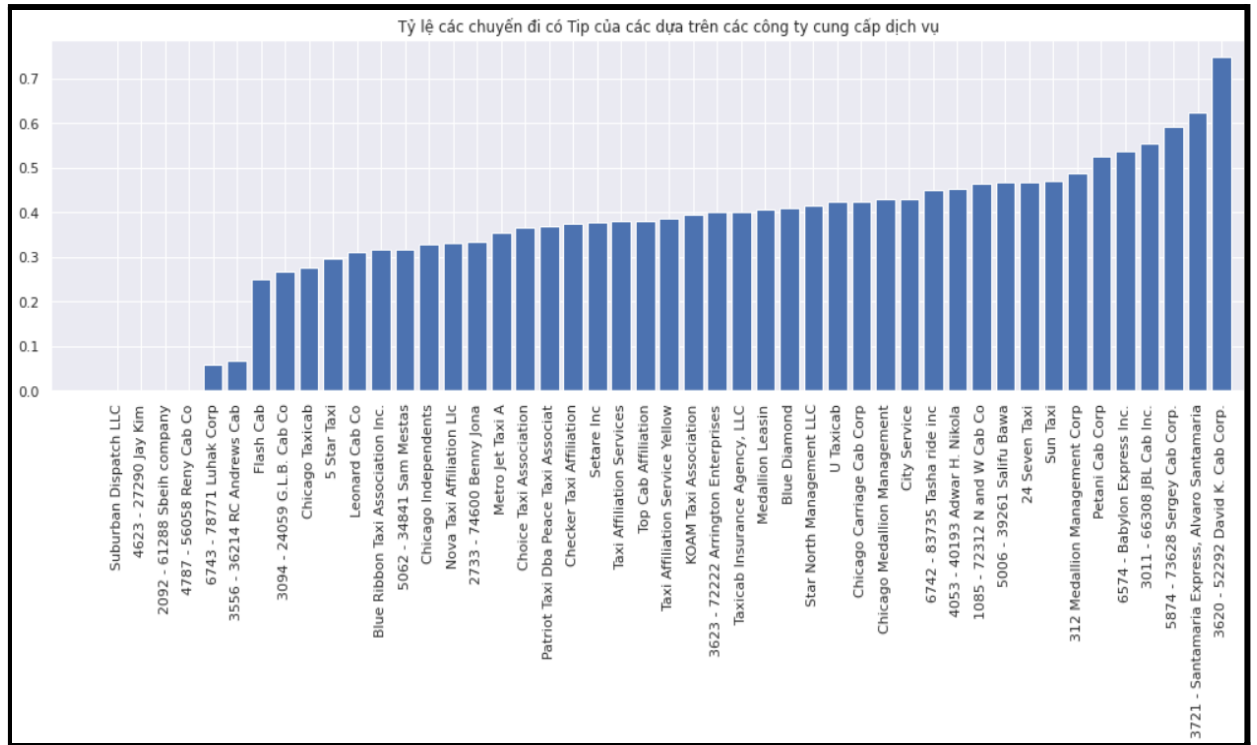


'Payment Type' có nhiều phương thức thanh toán không dùng tiền mặt: credit card, mobile, Prcard đúng với xu hướng chung trong việc phát triển thực tế là không sử dụng tiền mặt.

Phương thức thanh toán có sự chênh lệch lớn giữa các phương thức nên có lẽ thuộc tính này có sự tương quan lớn đến số tiền Tip nên sẽ giữ lại trong mô hình dự đoán.



Sự phân bố tiền Tip theo công ty đa số là tỉ lệ trên dưới 50% trên đa số các công ty nên thuộc tính này không có sự tương quan với tiền Tip nên sẽ không chọn thuộc tính này vào mô hình dự đoán.



2.4. Tiền xử lý dữ liệu và trích chọn đặc trưng

Dữ liệu trống:

Bảng 2: Các cột có dữ liệu trống và số lượng.

Tên thuộc tính	Số dòng trống	Tên thuộc tính	Số dòng trống
Trip Seconds	28	Pickup Centroid Latitude	13031
Trip Miles	2	Pickup Centroid Longitude	13031
Pickup Census Tract	87901	Pickup Centroid Location	13031
Dropoff Census Tract	88842	Dropoff Centroid Latitude	17903
Pickup Community Area	13204	Dropoff Centroid Longitude	17903
Dropoff Community Area	18393	Dropoff Centroid Location	17903

Xử lý dữ liệu kiểu Datetime

Có 2 thuộc tính có kiểu dữ liệu datetime: 'Trip Start Timestamp' và 'Trip End Timestamp'. Từ 2 cột này ta sẽ tạo ra 4 cột mới 'start time', 'start daytime', 'end time', 'end daytime' đồng nghĩa với việc 2 cột Timestamp này sẽ bị loại bỏ. Lý do bỏ đi các

thông tin ngày tháng vì trong bộ dữ liệu chỉ nằm trong một tháng nên không có nhiều ý nghĩa.

<i>Trip Start Timestamp</i>	<i>Trip End Timestamp</i>		<i>start time</i>	<i>start daytime</i>	<i>end time</i>	<i>end daytime</i>
01/01/2020 01:00:00 AM	01/02/2020 02:30:00 PM	→	01:00:00	morning	14:30:00	night
01/01/2020 02:30:00 AM	01/02/2020 12:15:00 PM		02:30:00	morning	12:15:00	night
01/01/2020 03:30:00 AM	01/01/2020 08:30:00 PM		03:30:00	morning	20:30:00	night

Xử lý thuộc tính '*Trip Seconds*'

Ta sử dụng hiệu của '*end time*' và '*start time*' để điền khuyết cho các giá ô trống.


<i>start time</i>	<i>end time</i>		<i>Trip Seconds</i>
01:00:00	14:30:00	→	48600.0
02:30:00	12:15:00		35100.0

Xử lý thuộc tính '*Trip Miles*'

Trong bộ dữ liệu có những điểm dữ liệu có '*Trip Seconds*' lớn hơn 0 nhưng lại có '*Trip Miles*' bằng 0 -> cần phải xử lý trường hợp này cùng với trường hợp khuyết giá trị.

<i>Pickup Centroid Location</i>	<i>Dropoff Centroid Location</i>	<i>Trip Seconds</i>	<i>Trip Miles</i>
P (-87.6333080367 41.899602111)	P (-87.667569312 41.8502663663)	1066.0	NaN
P (-87.6327464887 41.8809944707)	P (-87.6197106717 41.8950334495)	16972.0	NaN
P (-87.6559981815 41.9442266014)	P (-87.6333080367 41.899602111)	1200.0	0.0

Các điểm dữ liệu này được xử lý bằng cách tính khoảng cách Euclidean dựa vào tọa độ địa lý của nơi bắt đầu chuyến đi và nơi kết thúc chuyến đi, sau đó chuyển độ đo khoảng cách này về đơn vị dặm (mile) [1].

<i>Trip Miles</i>		<i>Trip Miles</i>
NaN		3.84
NaN		1.18
0.0		3.3

Xử lý '*Pickup Community Area*' và '*Dropoff Community Area*'

Như đã tìm được ở phần EDA, có mối liên hệ giữa Census Tract với Community Area -> giải pháp điển hình là tìm một bộ dữ liệu hỗ trợ [2] việc ánh xạ từ Census Tract sang Community Area.

<i>Census Tract</i>	<i>Community Area</i>
17031010701	1
17031010501	1
17031020301	2

Với những trường hợp không thể tìm được giá trị ánh xạ cho điểm dữ liệu thì sẽ loại bỏ điểm dữ liệu khỏi tập dữ liệu.

Lựa chọn các đặc trưng

Sau các bước xử lý bên trên ta đã tạo ra 4 đặc trưng mới ('*start time*', '*start daytime*', '*end time*', '*end daytime*') từ hai đặc trưng '*start time*' và '*end time*' ta đã tận dụng để điền khuyết và chỉnh sửa các giá trị cho cột '*Trip Seconds*'. Đồng thời ta sẽ 2 cột '*Trip Start Timestamp*' và '*Trip End Timestamp*' sẽ xoá bỏ và 2 cột mới '*start time*' và '*end time*' sau khi được sử dụng để điều chỉnh giá trị trong cột '*Trip Seconds*' cũng sẽ bị loại bỏ.

Các đặc trưng '*Pickup Census Tract*' và '*Dropoff Census Tract*' và '*Pickup Community Area*' và '*Dropoff Community Area*' có liên hệ với nhau nên ta chỉ giữ lại các Community Area vì các giá trị này có miền giá trị nhỏ hơn (1-77) nên việc không gian lưu trữ sẽ nhỏ hơn.

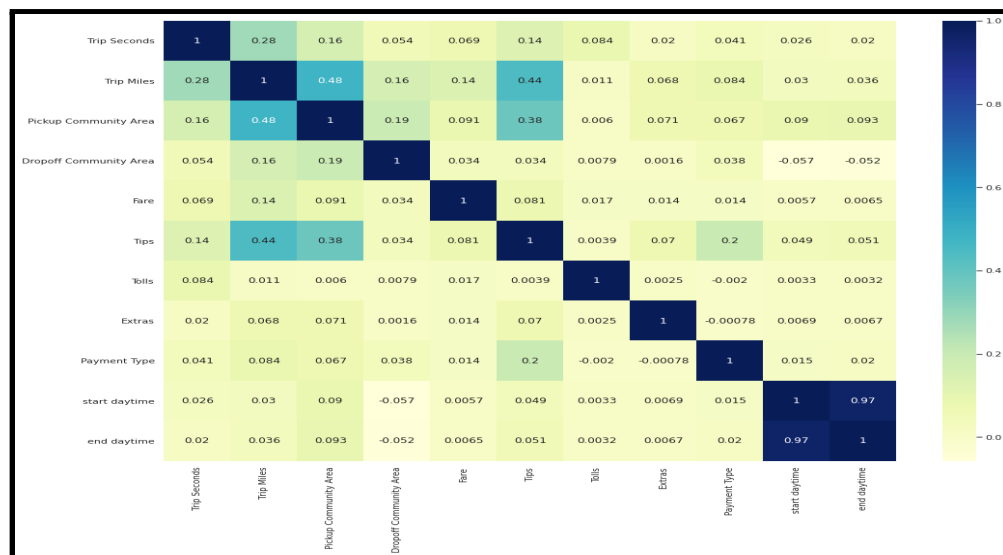
Các trường *'Pickup Centroid Latitude'*, *'Pickup Centroid Longitude'*, *'Pickup Centroid Location'*, *'Dropoff Centroid Latitude'*, *'Dropoff Centroid Longitude'*, *'Dropoff Centroid Location'* sẽ bị loại bỏ vì chỉ giúp ta biểu diễn trực quan các địa điểm và điền giá trị khuyết cũng như điều chỉnh giá trị của cột *'Trip Miles'*.

Ta cũng sẽ loại bỏ thêm cột *'Trip Total'* vì thuộc tính này được tính dựa vào tổng của *'Fare'*, *'Tolls'*, *'Extras'* và *'Tips'*.

Cùng với kết quả phân tích từ phần Phân tích thăm dò ta cũng nên bỏ đi cột *'Company'*.

Vậy ta còn lại tổng cộng là 10 biến độc lập và 1 biến mục tiêu: *'Trip Seconds'*, *'Trip Miles'*, *'Pickup Community Area'*, *'Dropoff Community Area'*, *'Fare'*, *'Tolls'*, *'Extras'*, *'Payment Type'*, *'start daytime'*, *'end daytime'* và *'Tips'*. Ta tiến hành phân tích tương quan giữa các thuộc tính.

Sau khi vẽ bảng tương quan giữa các thuộc tính ta loại bỏ thêm thuộc tính *'Tolls'* và *'end daytime'* vì *'Tolls'* có tương quan Pearson với *'Tips'* quá thấp (3%), hai thuộc tính *'start daytime'* và *'end daytime'* có tương quan với nhau rất cao (97%) nên chỉ cần chọn một trong hai thuộc tính, nhóm quyết định chọn ngẫu nhiên *'start time'*.



2.5. Lựa chọn và huấn luyện mô hình

Nhóm sử dụng phương pháp vét cạn (tổ hợp tất cả các thuộc tính được lựa chọn vd: *<Trip Seconds>*, *<Trips Seconds, start daytime>*,...).

Các tổ hợp thuộc tính sẽ được sử dụng để huấn luyện các mô hình Linear Regression và Multiple Linear Regression và Polynomial Regression (từ bậc 2 đến bậc 5).

Bộ dữ liệu được chia thành 2 tập Train và Test (tỷ lệ 8-2) với hai loại là dữ liệu có outlier và không có outlier để đánh giá thêm về dữ liệu outlier có ảnh hưởng tới việc xây dựng mô hình?

Sử dụng các độ đo đánh giá: RMSE, R2, 4-Fold Validation và 5-Fold Validation để đánh giá mô hình .

2.6. Đánh giá mô hình

Để thuận tiện cho việc hiển thị tên đặc trưng, các tên đặc trưng trong các hình bên dưới sẽ được viết tắt. (ví dụ: ‘Fare’: F, ‘Pickup Community Area’: PCA, ‘Trip Miles’: TM).

Hình ảnh kết quả đánh giá trên tập dữ liệu chưa xử lý outliers

	Model	Feature_details	RMSE	R ² _train	R ² _test	4_Fold_Validation	5_Fold_Validation	Note
0	Multiple Linear Regression	[TM, 'F', 'PCA', 'DCA', 'PT']	2.124521	0.254302	0.291223	0.252505	0.254774	Test_Size = 0.2, Number_Feature = 5
1	Multiple Linear Regression	[TM, 'PCA', 'DCA', 'PT']	2.124147	0.254015	0.291472	0.252777	0.254750	Test_Size = 0.2, Number_Feature = 4
2	Multiple Linear Regression	[TM, 'F', 'PCA', 'DCA', 'PT', 'SD']	2.124230	0.254459	0.291417	0.252389	0.254745	Test_Size = 0.2, Number_Feature = 6
3	Multiple Linear Regression	[TM, 'PCA', 'DCA', 'PT', 'SD']	2.123862	0.254172	0.291662	0.252664	0.254720	Test_Size = 0.2, Number_Feature = 5
4	Multiple Linear Regression	[TS, TM, 'F', 'PCA', 'DCA', 'PT']	2.124852	0.254369	0.291002	0.252178	0.254228	Test_Size = 0.2, Number_Feature = 6

	Model	Feature_details	RMSE	R ² _train	R ² _test	4_Fold_Validation	5_Fold_Validation	Note
0	PolynomialFeatures	[F, 'PT']	1.182926	0.744581	0.780263	0.732633	0.733684	Test_Size = 0.2, Degree = 4, Number_Feature = 2
1	PolynomialFeatures	[F, 'PCA', 'PT']	1.235830	0.726472	0.760169	0.728137	0.729938	Test_Size = 0.2, Degree = 3, Number_Feature = 3
2	PolynomialFeatures	[F, 'PCA', 'PT']	1.170155	0.746315	0.784982	0.723116	0.728944	Test_Size = 0.2, Degree = 4, Number_Feature = 3
3	PolynomialFeatures	[F, 'DCA', 'PT']	1.252279	0.722129	0.753742	0.724216	0.725472	Test_Size = 0.2, Degree = 3, Number_Feature = 3
4	PolynomialFeatures	[F, 'PT']	1.250390	0.721849	0.754485	0.724082	0.725329	Test_Size = 0.2, Degree = 3, Number_Feature = 2
5	PolynomialFeatures	[TS, 'F', 'PCA', 'PT']	1.233431	0.727044	0.761099	0.721382	0.724598	Test_Size = 0.2, Degree = 3, Number_Feature = 4
6	PolynomialFeatures	[TS, 'F', 'PT']	1.245558	0.722823	0.756378	0.712413	0.714251	Test_Size = 0.2, Degree = 3, Number_Feature = 3
8	PolynomialFeatures	[TS, 'F', 'DCA', 'PT']	1.247897	0.723195	0.755462	0.703290	0.704504	Test_Size = 0.2, Degree = 3, Number_Feature = 4
9	PolynomialFeatures	[F, 'PCA', 'DCA', 'PT']	1.279389	0.727338	0.742964	0.697954	0.700462	Test_Size = 0.2, Degree = 3, Number_Feature = 4
10	PolynomialFeatures	[TS, 'F', 'PCA', 'DCA', 'PT']	1.266162	0.727860	0.748252	0.694238	0.699459	Test_Size = 0.2, Degree = 3, Number_Feature = 5

Hình ảnh kết quả đánh giá trên tập dữ liệu đã xử lý outliers

	Model	Feature_details	RMSE	R ² _train	R ² _test	4_Fold_Validation	5_Fold_Validation	Note
0	Multiple Linear Regression	[TS, TM, 'F', 'PCA', 'DCA', 'PT', 'SD']	2.132395	0.295444	0.287609	0.285145	0.288459	Test_Size = 0.2, Number_Feature = 7
1	Multiple Linear Regression	[TS, TM, 'PCA', 'DCA', 'PT', 'SD']	2.132490	0.295382	0.287546	0.285190	0.288427	Test_Size = 0.2, Number_Feature = 6
2	Multiple Linear Regression	[TS, TM, 'F', 'PCA', 'DCA', 'PT']	2.132839	0.295254	0.287312	0.285193	0.288419	Test_Size = 0.2, Number_Feature = 6
3	Multiple Linear Regression	[TS, TM, 'PCA', 'DCA', 'PT']	2.132933	0.295192	0.287250	0.285235	0.288388	Test_Size = 0.2, Number_Feature = 5
4	Multiple Linear Regression	[TM, 'F', 'PCA', 'DCA', 'PT', 'SD']	2.132698	0.295323	0.287406	0.285060	0.288338	Test_Size = 0.2, Number_Feature = 6

	Model	Feature_details	RMSE	R ² _train	R ² _test	4_Fold_Validation	5_Fold_Validation	Note
0	PolynomialFeatures	['TM', 'F', 'PCA', 'PT']	1.208721	0.777955	0.771106	0.734591	0.770480	Test_Size = 0.2, Degree = 4, Number_Feature = 4
1	PolynomialFeatures	['F', 'PT']	1.188790	0.775022	0.778592	0.757601	0.758269	Test_Size = 0.2, Degree = 4, Number_Feature = 2
2	PolynomialFeatures	['TM', 'F', 'PCA', 'PT']	1.224030	0.757853	0.765271	0.754153	0.756583	Test_Size = 0.2, Degree = 3, Number_Feature = 4
3	PolynomialFeatures	['TM', 'F', 'PCA', 'PT', 'SD']	1.223078	0.757981	0.765636	0.745370	0.755202	Test_Size = 0.2, Degree = 3, Number_Feature = 5
4	PolynomialFeatures	['F', 'PCA', 'PT']	1.230430	0.755341	0.762810	0.753045	0.754662	Test_Size = 0.2, Degree = 3, Number_Feature = 3
5	PolynomialFeatures	['TM', 'F', 'PT']	1.214457	0.775834	0.768928	0.751412	0.754644	Test_Size = 0.2, Degree = 4, Number_Feature = 3
6	PolynomialFeatures	['F', 'PCA', 'PT']	1.182921	0.776811	0.780773	0.747612	0.753809	Test_Size = 0.2, Degree = 4, Number_Feature = 3
7	PolynomialFeatures	['TM', 'F', 'PT']	1.234203	0.753355	0.761353	0.750805	0.751868	Test_Size = 0.2, Degree = 3, Number_Feature = 3
8	PolynomialFeatures	['TS', 'TM', 'F', 'PCA', 'PT']	1.222385	0.758450	0.765901	0.747554	0.751688	Test_Size = 0.2, Degree = 3, Number_Feature = 5
9	PolynomialFeatures	['TM', 'F', 'DCA', 'PT']	1.235925	0.754390	0.760686	0.750609	0.751081	Test_Size = 0.2, Degree = 3, Number_Feature = 4

Từ các kết quả thực nghiệm trên dữ liệu ta có thể rút ra một số nhận xét:

- + Các mô hình Linear cho kết quả chưa tốt bằng các mô hình Polynomial
- + Kết quả trên bộ đã xử lý outliers cho kết quả tốt hơn.
- + Các feature quan trọng xuất hiện ở hầu hết các mô hình tốt nhất là: *'Fare'*, *'Payment Type'*, *'Pickup Community Area'*.

3. KẾT LUẬN

Sau các bước thăm dò tiền xử lý và trích chọn đặc trưng đã được trình bày ở trên. Từ 23 thuộc tính sau khi thăm dò và trích chọn đặc trưng để chạy model là 8 thuộc tính.

Kết quả cuối cùng cho mô hình tốt nhất đạt giá trị là 0.77 trên 5 fold validation với 4 thuộc tính là *'Trip Miles'*, *'Fare'*, *'Pickup Community Area'*, *'Payment Type'*.

Thuộc tính quan trọng nhất thuộc bộ dữ liệu là *'Pickup Community Area'*, *'Payment Type'* và *'Fare'*.

Thuộc tính được thêm vào nhờ phân tích Timestamp là *'start daytime'* cũng cho 1 mô hình có kết quả tốt là 0.75.

Ta trả lời được câu hỏi đặt ra ở phần Lựa chọn và huấn luyện mô hình đó là các outlier trong dữ liệu có ảnh hưởng tới các mô hình.

TÀI LIỆU THAM KHẢO

- [1] python.org. *Haversine*, <https://pypi.org/project/haversine/>. Accessed 19 12 2020.
- [2] Chicago Portal. *Chicago Data Portal*,
<https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Census-Tracts-2010/5jrd-6zik>. Accessed 21 12 2020.

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Trọng Ân	
2	Dương Văn Bình	