Phân tích và xây dựng mô hình dự đoán thời gian chuyến đi của taxi

Hồ Ngọc Mai[1,2,3], Lê Ngọc Thiên Phúc[1,2,3], Trần Đại Hiển[1,2,3], Võ Hoàng Thảo Phương[1,2,3], and Đỗ Trọng Hợp[1,2,4]

[1] Trường Đại học Công nghệ Thông tin, Thành phố Hồ Chí Minh, Việt Nam
[2] Đại học Quốc gia, Thành phố Hồ Chí Minh, Việt Nam
[3] {22520839,22521117,22520426,22521171}@gm.uit.edu.vn
[4] {hopdt@uit.edu.vn}

Tóm tắt nội dung Trong hệ thống giao thông đô thị, dự đoán chính xác thời gian chuyến đi taxi là yếu tố quan trọng để nâng cao hiệu quả dịch vụ và sự hài lòng của khách hàng. Nghiên cứu này khám phá việc ứng dụng các mô hình học máy để dự đoán thời lượng chuyến đi taxi tại Chicago. Bằng cách sử dụng bộ dữ liệu chuyến đi taxi tại Chicago kết hợp với các đặc trưng về thời tiết và tắc nghẽn giao thông Chicago, nghiên cứu này chứng minh tính hiệu quả của các kỹ thuật dự đoán trong việc ước tính thời gian di chuyển, từ đó hỗ trợ quản lý giao thông đô thị và tối ưu hóa tài nguyên. Nghiên cứu tập trung vào việc phân tích và xây dựng mô hình dự đoán thời gian chuyến đi taxi bằng cách áp dụng các mô hình học máy Linear Regression và Ridge Regression, với độ đo đánh giá đánh giá RMSLE (Root Mean Squared Logarithmic Error) và kỹ thuật cross-validation 10-fold để đảm bảo tính chính xác và độ tin cậy của mô hình. Bên cạnh đó, nhóm còn kiểm định tác động của điều kiện thời tiết và tắc nghẽn giao thông lên thời lượng chuyến đi taxi nhằm nâng cao hiệu suất mô hình.

Keywords: Thời gian chuyến đi của taxi · Linear Regression · Ridge Regression · RMSLE · cross-validation 10-fold

1 Giới thiệu

Với sự phát triển mạnh mẽ của đô thị hóa và công nghệ, nhu cầu sử dụng dịch vụ taxi ngày càng tăng. Việc dự đoán chính xác thời gian chuyến đi của taxi không chỉ giúp cải thiện chất lượng dịch vụ mà còn nâng cao hiệu quả hoạt động trong việc quản lý giao thông đô thị và tối ưu hóa tài nguyên. Chính vì những điều trên, nhóm đã đặt ra mục tiêu nghiên cứu chi tiết về những yếu tố ảnh hưởng đến thời gian chuyến đi taxi và xây dựng mô hình dự đoán thời gian chuyến đi của taxi thích hợp, để giúp các doanh nghiệp và cơ quan quản lý đô thị có thể điều hành giao thông đô thị ngày càng hiệu quả hơn. Bài toán phân tích và xây dựng mô hình dự đoán thời gian chuyến đi của taxi tại Chicago là bài toán mà nhóm cần xử lý với dữ liệu đầu vào là dữ liệu về chuyến đi taxi tại Chicago bao gồm những đặc điểm như vị trí điểm đón và điểm trả khách, thời

điểm bắt đầu chuyến đi, khoảng cách chuyến đi,..... Bằng cách áp dụng các mô hình học máy như Linear Regression và Ridge Regression với những dữ liệu đầu vào như trên, nhóm đã xây dựng mô hình dự đoán với dữ liệu đầu ra là thời gian chuyến đi của taxi tại Chicago. Mô hình được đánh giá thông qua phương pháp đánh giá chủ yếu RMSLE (Root Mean Squared Logarithmic Error) với kỹ thuật cross-validation 10-fold nhằm đảm bảo tính chính xác và độ tin cậy của mô hình.

Bên cạnh đó, nhóm còn kiểm định sự tác động của các yếu tố ảnh hưởng đến thời gian chuyến đi của taxi như yếu tố thời tiết như nhiệt độ, lượng mưa, độ ẩm không khí, tốc độ gió và yếu tố tắc nghẽn như mật độ giao thông tại thời điểm chuyến đi để cải thiện và nâng cao hiệu suất mô hình. Nghiên cứu đóng góp một mô hình dự đoán về thời gian chuyến đi của taxi dựa vào bộ dữ liệu hoàn chỉnh giúp cải thiện dự báo và quản lý hoạt động vận chuyển trong thành phố. Bằng cách phân tích và so sánh các mô hình, chúng tôi mong muốn cung cấp những hiểu biết mới mẻ về ảnh hưởng của các yếu tố này đến tính chính xác của dự đoán, từ đó giúp cải thiện hiệu quả vận hành trong ngành taxi và hệ thống giao thông đô thị. Bài viết này trình bày phân tích các yếu tố ảnh hưởng đến thời lượng chuyến đi bằng taxi và phác thảo phương pháp được sử dụng để xây dựng và đánh giá các mô hình dự đoán, góp phần thúc đẩy phân tích giao thông vận tải.

2 Bộ dữ liệu

Bộ dữ liệu mà nhóm sử dụng trong nghiên cứu là bộ dữ liệu toàn diện và đa dạng, được nhóm thu thập từ trang data.cityofchicago.org, một nguồn thông tin đáng tin cậy và đầy đủ về các chuyến đi taxi trong khoảng thời gian từ năm 2013 đến 2023. Trang web này cung cấp dữ liệu chất lượng cao, được cập nhật và kiểm duyệt kỹ lưỡng, đảm bảo tính chính xác và độ tin cậy cho các nghiên cứu và phân tích dữ liệu liên quan đến giao thông đô thị. Bộ dữ liệu ban đầu của nhóm bao gồm 3,783,730 dòng và 23 cột thuộc tính như Bảng 1.

Dữ liệu thời tiết được nhóm thu thập từ kaggle.com, một nền tảng nổi tiếng cung cấp các bộ dữ liệu chất lượng cao cho các nhà nghiên cứu và chuyên gia phân tích dữ liệu. Các bộ dữ liệu trên Kaggle thường được kiểm duyệt và đánh giá kỹ lưỡng bởi cộng đồng, đảm bảo tính chính xác và đáng tin cậy cho các nghiên cứu. Nguồn dữ liệu này cung cấp các thông tin chi tiết về thời tiết tại Chicago như nhiệt độ, lượng mưa, độ ẩm, và tốc độ gió. Bộ dữ liệu này được nhóm sử dụng nhằm mục đích trích xuất một số đặc trưng về thời tiết, sau đó bổ sung vào bộ dữ liệu ban đầu để khám phá mối tương quan giữa các yếu tố này và thời lượng chuyến đi taxi. Từ đó giúp nhóm nghiên cứu để cải thiện và nâng cao hiệu suất mô hình dự đoán thời gian chuyến đi taxi. Bộ dữ liệu thời tiết bao gồm 24,108 dòng và 10 cột thuộc tính như Bảng 2.

Ngoài việc thêm những đặc trưng về thời tiết, đặc trưng về tắc nghẽn giao thông cũng được nhóm bổ sung cho bộ dữ liệu này nhằm phân tích những ảnh

Bảng 1. Mô tả thuộc tính dữ liệu chuyến đi taxi ở Chicago.

Tên thuộc tính	**Mô tả thuộc tính**
Trip ID	Mã định danh duy nhất cho mỗi chuyến đi taxi.
Taxi ID	Mã định danh duy nhất cho mỗi taxi.
Trip Start Timestamp	Thời gian bắt đầu chuyến đi.
Trip End Timestamp	Thời gian kết thúc chuyến đi.
Trip Seconds	Thời gian chuyến đi tính bằng giây.
Trip Miles	Khoảng cách chuyến đi tính bằng dặm.
Pickup Census Tract	Mã vùng thống kê dân số nơi đón khách.
Dropoff Census Tract	Mã vùng thống kê dân số nơi trả khách.
Pickup Community Area	Khu vực cộng đồng nơi đón khách.
Dropoff Community Area	Khu vực cộng đồng nơi trả khách.
Fare	Tiền cước của chuyến đi.
Tips	Tiền tip mà khách hàng đưa cho tài xế.
Tolls	Tiền thu phí đường bộ của chuyến đi.
Extras	Các khoản phụ thu khác.
Trip Total	Tổng số tiền của chuyến đi (cước + tip + phí).
Payment Type	Hình thức thanh toán (tiền mặt, thẻ,...).
Company	Tên công ty taxi.
Pickup Centroid Latitude	Vĩ độ của điểm đón khách.
Pickup Centroid Longitude	Kinh độ của điểm đón khách.
Pickup Centroid Location	Vị trí điểm đón khách (kinh độ và vĩ độ).
Dropoff Centroid Latitude	Vĩ độ của điểm trả khách.
Dropoff Centroid Longitude	Kinh độ của điểm trả khách.
Dropoff Centroid Location	Vị trí điểm trả khách (kinh độ và vĩ độ).

Bảng 2. Mô tả thuộc tính dữ liệu thời tiết ở Chicago.

Tên thuộc tính	**Mô tả thuộc tính**
Year	Năm của dữ liệu được ghi lại.
Month	Tháng của dữ liệu được ghi.
Day	Ngày ghi dữ liệu.
Hour	Giờ của dữ liệu được ghi.
Temperature	Nhiệt độ tại thời điểm nhất định.
Precipitation	Lượng mưa tại thời điểm nhất định.
Humidity	Độ ẩm tại thời điểm nhất định.
Wind speed	Tốc độ gió tại thời điểm nhất định.
Atmosphere press	Áp suất khí quyển tại thời điểm nhất định.
REF	Mã tham chiếu cho dữ liệu.

hưởng của chúng đối với thời gian chuyến đi taxi. Bộ dữ liệu này là cơ sở để nhóm phát triển các mô hình học máy nhằm dự đoán thời gian chuyến taxi với độ chính xác và độ tin cậy cao.

3 Tiền xử lý dữ liệu

3.1 Làm sạch dữ liệu

Trong phần này, quy trình tiền xử lý dữ liệu được áp dụng cho tập dữ liệu chứa thông tin về chuyến đi của taxi ở Chicago trong năm 2023 gồm 23 cột thuộc tính và 3,783,730 dòng dữ liệu. Mục tiêu là tạo ra một tập dữ liệu rõ ràng, phong phú và phù hợp để khai thác.

Làm sạch dữ liệu chuyến đi taxi

- **Missing Values**: Đối với các dòng dữ liệu bị khuyết, thực hiện việc kiểm tra và nhận thấy lượng dữ liệu thiếu không đáng kể. Nhóm tiến hành loại bỏ các dòng dữ liệu đó.
- **Outlier Removal – Trip Seconds**: Xác định và loại bỏ các dữ liệu nhiễu trên thuộc tính Trip Seconds bằng cách sử dụng độ lệch trung bình và độ lệch chuẩn (mean and standard deciation).
- **Coordinate Cleanup**: Xác định tọa độ thật của biên giới Chicago, loại bỏ dữ liệu có địa điểm đón, trả khách từ các vùng ngoài biên giới Chicago.
- **Feature Engineering**: Xóa bỏ các chuyến đi có điểm đón, trả khách trùng nhau. Tính khoảng cách giữa điểm đón và trả khách bằng công thức khoảng cách Manhattan.
- **Feature Extraction**: Trích xuất các đặc trưng như tháng, giờ , tuần trong năm, ngày trong tuần và ngày trong tháng từ cột ngày giờ.
- **Filtering**: Lọc lại dữ liệu dựa trên các tham số liên quan. Đặt thời lượng di chuyển tối thiểu bằng 120 giây. Đặt ngưỡng khoảng cách tối đa và tối thiểu dựa trên khoảng cách thực tế.

Kết hợp các dữ liệu thời tiết và tắc nghẽn giao thông

Nhóm tiến hành chọn các cột thuộc tính có liên quan từ dữ liệu thời tiết (temperature, precipitation, humidity, wind speed) để thêm vào dữ liệu chuyến đi taxi tương ứng dựa trên thời gian chạy của từng chuyến taxi.
Để đánh giá thêm yếu tố tác động lên thời gian di chuyển của taxi, nhóm quyết định thêm thuộc tính mới là giờ cao điểm (is-rush-hour) dựa trên thời gian khi đón khách.
Quá trình làm sạch dữ liệu này đã tạo ra một tập dữ liệu có cấu trúc hoàn chỉnh, chứa nhiều thông tin hữu ích trong việc phân tích và xây dựng mô hình sắp tới.

3.2 Khai phá dữ liệu

Nhóm đã tiến hành phân tích số lượng chuyến đi của taxi theo giờ, ngày và ngày trong tuần, cũng như tác động của dữ liệu thời tiết và tắc nghẽn giao thông đến thời gian di chuyển của taxi tại Chicago nhằm xác định các yếu tố ảnh hưởng chính ảnh hưởng giúp xây dựng mô hình dự đoán thời gian di chuyển của taxi một cách chính xác, nhằm nâng cao hiệu suất mô hình.

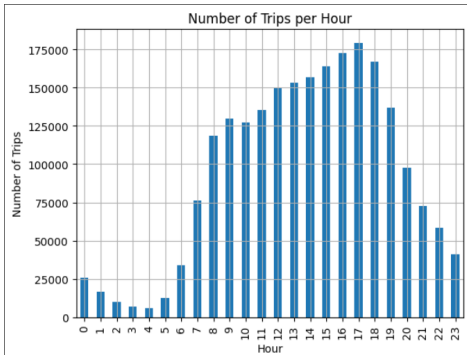
```
RangeIndex: 3783730 entries, 0 to 3783729
Data columns (total 23 columns):
 #   Column                      Dtype
---  ------                      -----
 0   Trip ID                     object
 1   Taxi ID                     object
 2   Trip Start Timestamp        object
 3   Trip End Timestamp          object
 4   Trip Seconds                float64
 5   Trip Miles                  float64
 6   Pickup Census Tract         float64
 7   Dropoff Census Tract        float64
 8   Pickup Community Area       float64
 9   Dropoff Community Area      float64
 10  Fare                        float64
 11  Tips                        float64
 12  Tolls                       float64
 13  Extras                      float64
 14  Trip Total                  float64
 15  Payment Type                object
 16  Company                     object
 17  Pickup Centroid Latitude    float64
 18  Pickup Centroid Longitude   float64
 19  Pickup Centroid Location    object
 20  Dropoff Centroid Latitude   float64
 21  Dropoff Centroid Longitude  float64
 22  Dropoff Centroid  Location  object
dtypes: float64(15), object(8)
```

Dữ liệu ban đầu

```
RangeIndex: 2253047 entries, 0 to 2253046
Data columns (total 19 columns):
 #   Column                  Dtype
---  ------                  -----
 0   pickup_datetime         object
 1   Trip Seconds            float64
 2   Company                 int64
 3   Pickup Community Area   float64
 4   Dropoff Community Area  float64
 5   pickup_latitude         float64
 6   pickup_longitude        float64
 7   dropoff_latitude        float64
 8   dropoff_longitude       float64
 9   month                   int64
 10  hour                    int64
 11  day_of_week             int64
 12  day                     int64
 13  distance                float64
 14  tempearture             float64
 15  precipitation           float64
 16  humidity                float64
 17  wind_speed              float64
 18  is_rush_hour            int64
dtypes: float64(12), int64(6), object(1)
```

Dữ liệu sau khi tiền xử lý

Hình 1. Bộ dữ liệu trước và sau khi tiền xử lý



Hình 2. Sự phân bố số lượng chuyến đi theo giờ

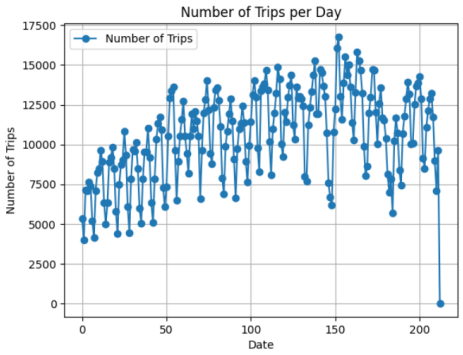
Phân tích số lượng chuyến đi theo giờ, theo ngày và theo ngày trong tuần

Dựa vào hình 2, ta thấy số lượng chuyến đi tập trung trung nhiều nhất từ khoảng 16-18 giờ, cho thấy nhu cầu sử dụng dịch vụ taxi tăng cao vào buổi chiều và chiều tối.

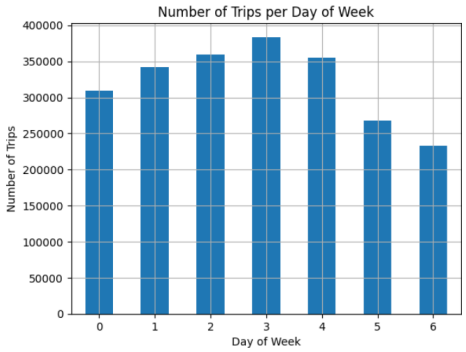
Theo hình 3, nhóm đưa ra nhận định rằng số lượng chuyến đi taxi tập trung nhiều nhất vào mùa hè, phản ánh nhu cầu sử dụng taxi tăng cao trong thời gian này.

Hình 4 cho thấy số lượng chuyến đi tập trung nhiều nhất từ thứ 4 đến thứ 6 hàng tuần, cho thấy nhu cầu sử dụng taxi tăng cao vào giữa tuần.

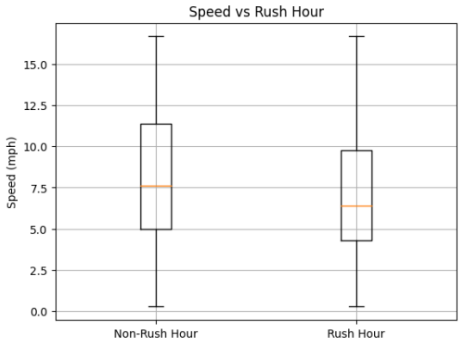
Nhìn chung, số lượng chuyến đi có xu hướng thay đổi rõ rệt theo giờ và theo ngày, phản ánh nhu cầu sử dụng dịch vụ taxi thay đổi theo thời gian.



Hình 3. Sự phân bố số lượng chuyến đi theo ngày



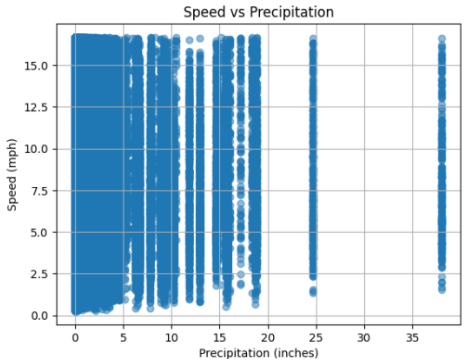
Hình 4. Sự phân bố số lượng chuyến đi theo ngày trong tuần



Hình 5. Tác động của giờ cao điểm đến tốc độ chuyến đi của taxi

Tác động của tắc nghẽn giao thông đến thời gian di chuyển taxi

Nhóm sử dụng biểu đồ hộp biểu diễn ảnh hưởng của tắc nghẽn giao thông trong khoảng thời gian cao điểm đến tốc độ xe. Dữ liệu được nhóm lại theo giờ cao điểm (6-8 và 16-18) và giờ thấp điểm (các giờ còn lại).
Theo hình 5, tốc độ xe trong giờ cao điểm thấp hơn rõ rệt so với giờ thấp điểm, dẫn đến thời gian di chuyển của taxi trong giờ cao điểm cao hơn nhiều.
Từ nhận định trên, ta thấy dữ liệu tắc nghẽn giao thông có ảnh hưởng rõ rệt đến thời gian di chuyển của taxi, thể hiện qua sự biến đổi trong thời gian di chuyển vào các khung giờ cao điểm trong ngày.



Hình 6. Sự tác động của lượng mưa đến tốc độ di chuyển của taxi

Tác động của dữ liệu thời tiết đến thời gian di chuyển của taxi

Từ hình 6, nhóm nhận định rằng lượng mưa có tác động lớn đến tốc độ di chuyển của taxi. Các yếu tố thời tiết như nhiệt độ, độ ẩm, tốc độ gió, và áp suất khí quyển cũng ảnh hưởng ít nhiều đến thời gian di chuyển.
Từ nhận định trên, nhóm thấy rằng các yếu tố thời tiết như nhiệt độ, lượng mưa, độ ẩm, tốc độ gió, và áp suất khí quyển có tác động đến thời gian di chuyển của taxi.
Những nhận định trên cho thấy, việc kết hợp dữ liệu ban đầu với dữ liệu thời tiết và tắc nghẽn giao thông là cần thiết để nâng cao hiệu suất mô hình dự đoán thời gian di chuyển của taxi tại Chicago.

4 Lý thuyết và mô hình

4.1 Mô hình máy học

Hồi quy tuyến tính (Linear Regression)

Hồi quy tuyến tính là một kỹ thuật phân tích dữ liệu dự đoán giá trị của dữ liệu không xác định bằng cách sử dụng một giá trị dữ liệu liên quan và đã biết khác. Nó mô hình toán học biến không xác định hoặc phụ thuộc và biến đã biết hoặc độc lập như một phương trình tuyến tính.[1]

Hồi quy Ridge (Ridge Regression)

Hồi quy Ridge là một biến thể của hồi quy tuyến tính, cho phép hạn chế hệ số của một mô hình. Các nhà khoa học dữ liệu có thể điều chỉnh một hệ số phạt, bù đắp cho ảnh hưởng của hệ số đó trong khi lập mô hình kết quả. Hệ số thông số có thể được triệt tiêu xuống gần bằng 0 trong mô hình hồi quy Ridge. Mô hình hồi quy Ridge là một phương pháp hồi quy tuyến tính được sử dụng khi dữ liệu có vấn đề đa cộng tuyến. Nó giúp cải thiện độ chính xác mô hình và ngăn chặn hiện tượng overfitting.[2]

4.2 Hướng tiếp cận

ANOVA

One-way ANOVA so sánh trung bình của hai hoặc nhiều nhóm cho một biến phụ thuộc và được sử dụng khi có hơn hai nhóm (khác với t-test). ANOVA sử dụng F-statistic, là tỷ lệ giữa phương sai giữa các nhóm và phương sai trong nhóm, tập trung vào sự khác biệt về phương sai để phân tích sự khác biệt về trung bình nhóm[3].
Công thức của ANOVA được nhóm trình bày ở hình 7. Trong đó:

$$\mathrm{F} = \frac{Intergroup\ variance}{Intragroup\ variance} = \frac{\sum_{i=1}^{K} n_i \left(\bar{Y}_i - \bar{Y}\right)^2 / \left(K-1\right)}{\sum_{ij=1}^{n} \left(Y_{ij} - \bar{Y}_i\right)^2 / \left(N-K\right)}$$

Hình 7. Công thức của ANOVA

$\bar{Y}i$ là trung bình của nhóm i; ni là số quan sát của nhóm i; $\bar{Y}$ là trung bình tổng thể; K là số nhóm; Yij là giá trị quan sát thứ j của nhóm i; và N là tổng số giá trị quan sát. F-statistic là tỷ lệ giữa tổng bình phương trung bình giữa các nhóm và tổng bình phương trung bình trong nhóm.

Post-hoc test

Kiểm định hậu kiểm (post-hoc) được thực hiện vì các kết luận từ ANOVA có giới hạn. Khi ANOVA bác bỏ giả thuyết không (rằng trung bình của các nhóm là như nhau), điều này chỉ cho biết ít nhất một nhóm có sự khác biệt mà không chỉ rõ nhóm nào có sự khác biệt. Do đó, ta tiến hành hậu kiểm so sánh các cặp nhóm khác nhau để xác định sự khác biệt cụ thể[5]. Trong đồ án này, nhóm sẽ dùng Tukey HSD.

Kiểm định Tukey HSD (honestly significant difference) được dùng để xác định rằng mối quan hệ giữa hai tập dữ liệu có ý nghĩa thống kê hay không. Tukey so sánh sự khác biệt giữa các trung bình thay vì các giá trị riêng lẻ. Giá trị của kiểm định được tính bằng cách chia sự khác biệt tuyệt đối giữa các cặp trung bình cho sai số chuẩn (SE) của trung bình, được xác định trước từ one-way ANOVA. SE là căn bậc hai của phương sai chia cho kích thước mẫu.[5]

Kiểm định Tukey so sánh từng cặp trung bình tổng thể, để đo lường, ta sử dụng một biến gọi là HSD được tính với công thức như sau:

$$HSD = \frac{M_i - M_j}{\sqrt{\frac{MS_w}{n_h}}}$$

Hình 8. Công thức HSD

Trong đó:

- Mi – Mj là sự khác biệt giữa cặp trung bình. (với điều kiện Mi lớn hơn Mj)
- MSw là Giá trị trung bình bình phương trong nhóm, và N là số lượng trong nhóm

Các bước thực hiện:

- **Bước 1:** Thực hiện kiểm định ANOVA. Giả sử giá trị F và p-value có ý nghĩa; tiến hành kiểm định hậu kiểm.
- **Bước 2:** Chọn hai mẫu trung bình từ kết quả ANOVA. Chú ý các thông tin sau:
  - Trung bình
  - Giá trị trung bình bình phương trong nhóm
  - Số lượng trong mỗi nhóm
  - Độ tự do (degrees of freedom) trong nhóm
- **Bước 3:** Áp dụng phương pháp Tukey HSD sử dụng công thức như trên.
- **Bước 4:** Xác định hệ số HSD trong bảng giá trị đối chiếu của Tukey
- **Bước 5:** So sánh các giá trị thu được ở bước 3 và bước 4. Nếu giá trị tính toán từ bước 3 lớn hơn giá trị thu được từ bước 4 thì có thể kết luận rằng có sự khác nhau giữa trung bình cặp nhóm đó.

Độ đo đánh giá: Root Mean Squared Logarithmic Error (RMSLE)

Root Mean Squared Logarithmic Error (RMSLE) là một chỉ số được sử dụng để đo độ chính xác của một mô hình so với dữ liệu thực tế. RMSLE là phiên bản của RMSE áp dụng cho các giá trị được lấy logarithm trước khi tính toán, hữu ích hơn khi làm việc với dữ liệu có phân phối không đều hoặc có sự biến đổi lớn. [6]

RMSLE thể hiện sự khác biệt trung bình giữa giá trị dự đoán và giá trị thực tế. Nếu RMSLE càng thấp, mô hình dự đoán sẽ càng chính xác.

5 Thực nghiệm và phân tích kết quả

5.1 Thực nghiệm

Sau quá trình tiền xử lý dữ liệu, nhóm tiến hành thực nghiệm trên bốn tập dữ liệu:

- Tập dữ liệu gốc
- Tập dữ liệu gốc kết hợp với yếu tố thời tiết
- Tập dữ liệu gốc kết hợp với yếu tố tắc nghẽn
- Tập dữ liệu gốc kết hợp với yếu tố thời tiết và yếu tố tắc nghẽn

Nhóm sử dụng hai mô hình học máy: Linear Regression và Ridge Regression để chạy trên bốn tập dữ liệu với độ đo RMSLE.

Trong từng lần chạy trên mỗi mô hình, tập dữ liệu sẽ được chia nhỏ thành 10 phần tương ứng với 10 tập train/test khác nhau (cross-validation 10-fold). Sau đó, nhóm sẽ thu được 10 kết quả đối với mỗi mô hình trên mỗi tập dữ liệu. Cuối cùng, ta sẽ tính trung bình kết quả của hai mô hình để có một kết quả duy nhất ứng với từng tập dữ liệu.

5.2 Phân tích kết quả

	original_dataset	original_dataset+weather_features	original_dataset+congestion_features	original_dataset+weather+congestion_features
0	0.373556	0.371870	0.370051	0.367778
1	0.373504	0.371878	0.370164	0.367953
2	0.372320	0.370751	0.368973	0.366839
3	0.372931	0.371339	0.369448	0.367270
4	0.372736	0.371221	0.369252	0.367178
5	0.373085	0.371746	0.369647	0.367743
6	0.374029	0.372400	0.370666	0.368442
7	0.372071	0.370445	0.368778	0.366583
8	0.373079	0.371531	0.369824	0.367741
9	0.371679	0.370170	0.368190	0.366109

Hình 9. Kết quả dựa trên độ đo RMSLE

Sau quá trình tiến hành thực nghiệm, dựa vào hình 9, nhóm thu được kết quả RMSLE trung bình như sau:

- Dataset ban đầu: **0.3729**
- Dataset được thêm các yếu tố thời tiết: **0.3713**
- Dataset được thêm yếu tố tắc nghẽn: **0.3695**
- Dataset được thêm cả 2 yếu tố kể trên: **0.3674**

Dựa vào kết quả trung bình trên, có thể thấy rằng khi thêm các thuộc tính thời tiết và tắc nghẽn thì kết quả có cải thiện. Ngoài ra, khi kết hợp tất cả các yếu tố thêm vào thì cho ra kết quả tốt nhất. Để kiểm tra điều này xảy ra có phải là do ngẫu nhiên hay không, nhóm tiến hành các bước kiểm định đã đề cập như trên. Nhóm tiến hành phân tích ANOVA với các giả thiết sau:

- **H0:** Không có sự khác nhau giữa trung bình các nhóm
- **H1:** Có sự khác nhau giữa trung bình các nhóm

```
              sum_sq    df           F        PR(>F)
C(Model)    0.000171   3.0  111.517945  2.767761e-18
Residual    0.000018  36.0         NaN           NaN
ANOVA is significant, performing post-hoc tests...
```

Hình 10. Phân tích ANOVA

Lấy mức ý nghĩa **0.05** với hình 10, ta nhận thấy f-value là **2.767761e-18**. Vì vậy ta bác bỏ H0, cho thấy có sự khác nhau giữa trung bình các nhóm. Tiếp theo, để kiểm định có sự khác nhau hay không của từng cặp trong nhóm, ta tiến hành Post-hoc analysis bằng Tukey HSD.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
original_dataset	original_dataset+congestion_features	-0.0034	0.0	-0.0043	-0.0025	True
original_dataset	original_dataset+weather+congestion_features	-0.0055	0.0	-0.0064	-0.0047	True
original_dataset	original_dataset+weather_features	-0.0016	0.0001	-0.0024	-0.0007	True
original_dataset+congestion_features	original_dataset+weather+congestion_features	-0.0021	0.0	-0.003	-0.0013	True
original_dataset+congestion_features	original_dataset+weather_features	0.0018	0.0	0.001	0.0027	True
original_dataset+weather+congestion_features	original_dataset+weather_features	0.004	0.0	0.0031	0.0048	True

Hình 11. Post-hoc analysis sử dụng Tukey HSD

Dựa vào hình trên với mức ý nghĩa 0.05, từng cặp trong các kết quả đều có p-value < 0.05. Điều đó chứng tỏ có sự khác nhau giữa từng cặp kết quả. Vì thế, dựa vào kết quả trung bình ta có thể xác nhận sau khi thêm các thuộc tính thời tiết và tắc nghẽn thì kết quả có cải thiện so với dataset ban đầu bởi chỉ số RMSLE trung bình thấp hơn (lần lượt là **0.3713** và **0.3695** so với **0.3729**). Ngoài ra, khi kết hợp tất cả các yếu tố thêm vào thì cho ra kết quả tốt nhất với RMSLE: **0.3674**.

6 Kết luận và hướng phát triển

Trong nghiên cứu này, nhóm đã thực hiện tìm hiểu, phân tích và xây dựng mô hình dự đoán thời gian di chuyển của taxi dựa trên bộ dữ liệu chuyến đi của taxi tại Chicago. Trong nghiên cứu này, nhóm đã đóng góp những phần sau:

Với bộ dữ liệu ban đầu còn tồn tại nhiều vấn đề, nhóm đã quyết định tiến hành các phương pháp làm sạch dữ liệu cũng như phân tích và khám phá dữ liệu (EDA). Quá trình làm sạch dữ liệu nhằm loại bỏ các giá trị thiếu, không hợp lệ, không cần thiết và lọc lại dữ liệu, đảm bảo tính nhất quán của dữ liệu. Sau đó, nhóm tiến hành thêm các cột thuộc tính có liên quan từ dữ liệu thời tiết như nhiệt độ, độ ẩm, lượng mưa và áp suất khí quyển và tạo cột mới đánh dấu giờ cao điểm vào tập dữ liệu. Cuối cùng, nhóm tiến hành phân tích và khám phá bộ dữ liệu vừa xây dựng nhằm hiểu rõ hơn về cấu trúc, đặc điểm của dữ liệu và đặc biệt là xác định những tác nhân chính ảnh hưởng đến thời gian di chuyển của taxi để nâng cao hiệu suất mô hình.

Sau đó, nhóm thực nghiệm trên các tập dữ liệu khác nhau: dữ liệu gốc, dữ liệu sau khi thêm các đặc trưng về thời tiết, dữ liệu sau khi thêm cột đánh dấu giờ cao

điểm và dữ liệu sau khi thêm cả hai yếu tố trên bằng cách áp dụng trên các mô hình Linear Regression và Ridge Regression kết hợp với kỹ thuật cross-validation 10-fold. Đối với mỗi tập dữ liệu, nhóm thực hiện tính toán độ đo đánh giá RMSLE (Root Mean Squared Logarithmic Error) để đánh giá hiệu quả mô hình và sử dụng phương pháp ANOVA để kiểm định mô hình dự đoán tốt hơn với dữ liệu nào ở mức ý nghĩa 5%.

Tuy nhiên, đồ án của nhóm vẫn còn tồn tại một số điểm hạn chế và còn nhiều phần cần bổ sung và cải thiện. Các mô hình học máy nhóm đang thực hiện vẫn còn đơn giản cũng như chưa điều chỉnh tham số tối ưu nên chưa cho kết quả tối ưu hoàn toàn. Chính vì những lý do trên, để cải thiện nghiên cứu trong tương lai , nhóm sẽ phát triển theo những hướng sau:

Thêm thuộc tính mới: Xem xét việc thêm các thuộc tính quan trọng khác, ảnh hưởng trực tiếp đến thời gian di chuyển của taxi như lưu lượng giao thông, sự kiện trong thành phố, hay điều kiện đường xá.

Thực hiện Feature Selection: Thực hiện các kỹ thuật chọn lọc thuộc tính để loại bỏ những cột không có nhiều ý nghĩa, giúp giảm kích thước dữ liệu và cải thiện hiệu quả mô hình.

Thử nghiệm các phương pháp tiền xử lý khác: Thử nghiệm các phương pháp như chuẩn hóa dữ liệu, biến đổi dữ liệu hoặc xử lý missing values theo nhiều cách khác nhau để tìm ra phương pháp tốt nhất.

Thử nghiệm trên các mô hình học sâu và mô hình dựa trên đồ thị: Áp dụng các mô hình tiên tiến như mạng nơ-ron hồi quy (RNN), LSTM, hoặc các mô hình Time Series như ARIMA để cải thiện kết quả dự đoán. Ngoài ra, mô hình dựa trên đồ thị có thể giúp khám phá các mối quan hệ phức tạp giữa các thuộc tính trong dữ liệu.

Bằng cách thực hiện những cải tiến này, nhóm có thể tối ưu hóa mô hình dự đoán, giúp đưa ra những kết quả chính xác hơn và tận dụng tốt hơn tiềm năng của dữ liệu.

Tài liệu

1. Amazon Web Service, `https://aws.amazon.com/vi/what-is/linear-regression/`, last accessed 2024/06/26
2. All Thing IT, `https://ai.atsit.in/vi/posts/1172656640/`, last accessed 2024/06/27
3. Kamje, `https://synapse.koreamed.org/articles/1156679`, last accessed 2024/06/25
4. Kim, T. Understanding one-way ANOVA using conceptual figures. *Korean Journal Of Anesthesiology*. **70**, 22-26 (2017)
5. Nanda, A., Mohapatra, B., Mahapatra, A., Mahapatra, A. & Mahapatra, A. Multiple comparison test by Tukey's honestly significant difference (HSD): Do the confident level control type I error. *International Journal Of Statistics And Applied Mathematics*. **6**, 59-65 (2021)
6. Medium, What's the Difference Between RMSE and RMSLE? `https://medium.com/analytics-vidhya/root-mean-square-log-error-rmse-vs-rmlse-935c6cc1802a`, last accessed 2024/06/25