TRƯỜNG ĐẠI HỌC ĐẠI NAM

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN TÍCH XU HƯỚNG DI CHUYỂN TAXI TẠI NEW YORK**

**Giảng viên: Lê Thị Thùy Trang**

**Sinh viên thực hiện:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TT | Mã sv | Họ và Tên | Ngày Sinh | Lớp |
| 1 | 1571020026 | Nguyễn Duy Chiến | 13-10-2003 | CNTT-1602 |

**Hà Nội, năm 2024**

TRƯỜNG ĐẠI HỌC ĐẠI NAM

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN TÍCH XU HƯỚNG DI CHUYỂN TAXI TẠI NEW YORK**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TT | Mã sv | Họ và Tên | Ngày Sinh | Điểm | |
| Bằng Số | Bằng Chữ |
| 1 | 1571020026 | Nguyễn Duy Chiến | 13-10-2003 |  |  |

|  |  |
| --- | --- |
| **CÁN BỘ CHẤM THI 1**  **Trần Quý Nam** | **CÁN BỘ CHẤM THI 2** |

### 

**Hà Nội, năm 2024**

**LỜI NÓI ĐẦU**

New York, với hệ thống giao thông phức tạp và nhu cầu di chuyển cao, luôn là thị trường trọng điểm của ngành taxi. Trong những năm qua, sự phát triển của công nghệ, xu hướng sử dụng phương tiện bền vững và chính sách quản lý đô thị đã tác động mạnh mẽ đến hoạt động kinh doanh taxi tại thành phố này.

Sự xuất hiện của taxi điện, trí tuệ nhân tạo trong tối ưu hóa hành trình, và các nền tảng gọi xe công nghệ như Uber, Lyft đã thay đổi đáng kể hành vi của người dùng. Đồng thời, chính quyền New York cũng đưa ra nhiều quy định mới nhằm kiểm soát khí thải, cải thiện dịch vụ và đảm bảo tính cạnh tranh lành mạnh trong ngành vận tải.

Báo cáo này sẽ tập trung phân tích các dữ liệu về số lượng chuyến đi, xu hướng giá cước, hành vi người dùng, cũng như tác động của các chính sách mới đến ngành taxi. Bằng cách kết hợp dữ liệu lớn và các phương pháp phân tích hiện đại, chúng tôi đưa ra những dự báo quan trọng về sự phát triển của ngành trong tương lai, từ đó giúp doanh nghiệp và nhà quản lý có những chiến lược phù hợp để thích nghi với sự thay đổi không ngừng của thị trường.

**MỤC LỤC**

[**CHƯƠNG 1. TỔNG QUAN CƠ SỞ LÝ THUYẾT LIÊN QUAN** 5](#_Toc193133376)

[**1.Giới thiệu** 5](#_Toc193133377)

[***1.1* Bối cảnh nghiên cứu** 5](#_Toc193133378)

[***1.2. Mục tiêu nghiên cứu*** 5](#_Toc193133379)

[**CHƯƠNG 2. MÔ TẢ TẬP DỮ LIỆU VÀ CÔNG NGHỆ SỬ DỤNG** 7](#_Toc193133380)

[***2.1.* Công nghệ sử dụng** 7](#_Toc193133381)

[***2.2* . Dữ liệu và tiền xử lý** 7](#_Toc193133382)

[**CHƯƠNG 3. KẾT QUẢ XỬ LÝ, PHÂN TÍCH DỮ LIỆU VÀ ỨNG DỤNG** 9](#_Toc193133383)

[***3.1.*Phân tích dữ liệu** 9](#_Toc193133384)

[3.2. Mô hình hóa - Dự đoán giá cước taxi 19](#_Toc193133385)

[**KẾT LUẬN** 23](#_Toc193133386)

[**4. Kết luận và đề xuất** 23](#_Toc193133387)

[**4.1 Kết quả đạt được** 24](#_Toc193133388)

[4.2. Hạn chế và hướng phát triển 24](#_Toc193133389)

**CHƯƠNG 1. TỔNG QUAN CƠ SỞ LÝ THUYẾT LIÊN QUAN**

***1*.Giới thiệu**

***1.1* Bối cảnh nghiên cứu**

Dịch vụ taxi là một phần không thể thiếu trong hệ thống giao thông tại New York, phục vụ hàng triệu lượt di chuyển mỗi ngày. Với sự phát triển của công nghệ và dữ liệu lớn, việc phân tích dữ liệu taxi không chỉ giúp hiểu rõ hơn về thói quen di chuyển của người dân mà còn mang lại nhiều lợi ích trong việc tối ưu hóa hoạt động vận tải.

Dữ liệu từ các chuyến taxi cung cấp nhiều thông tin quan trọng như thời gian di chuyển, điểm đón và trả khách, khoảng cách, chi phí và tiền tip. Việc khai thác những dữ liệu này có thể giúp dự báo nhu cầu sử dụng taxi theo thời gian, xác định khu vực có lượng khách cao, phân tích doanh thu cũng như đá­­nh giá mức độ hài lòng của khách hàng dựa trên tiền tip.

Trong bối cảnh dữ liệu taxi có khối lượng rất lớn, các phương pháp xử lý truyền thống gặp nhiều hạn chế về tốc độ và hiệu suất. Do đó, việc sử dụng các công cụ như Apache Spark giúp xử lý dữ liệu nhanh hơn và hiệu quả hơn, từ đó đưa ra các phân tích sâu sắc, hỗ trợ các doanh nghiệp vận tải tối ưu hóa dịch vụ, đồng thời giúp các cơ quan quản lý có cái nhìn tổng quan về tình hình giao thông trong thành phố.

***1.2. Mục tiêu nghiên cứu***

Mục tiêu chính của nghiên cứu này là khai thác dữ liệu taxi tại New York để hiểu rõ hơn về xu hướng di chuyển, doanh thu, và hành vi của khách hàng. Trọng tâm của phân tích bao gồm các khía cạnh quan trọng sau:

Thứ nhất, nghiên cứu tập trung vào việc phân tích xu hướng di chuyển taxi theo thời gian, bao gồm các khung giờ cao điểm, sự khác biệt giữa ngày trong tuần và cuối tuần, cũng như xu hướng theo từng tháng trong năm. Điều này giúp xác định các khoảng thời gian có nhu cầu cao, hỗ trợ các công ty taxi trong việc tối ưu hóa lịch trình hoạt động.

Thứ hai, nghiên cứu sẽ xác định các điểm đón phổ biến nhất và phân tích nhu cầu di chuyển giữa các khu vực trong thành phố. Bằng cách xác định những vị trí có mật độ đặt xe cao, các nhà quản lý dịch vụ taxi có thể điều chỉnh số lượng xe phù hợp, giảm thời gian chờ và cải thiện chất lượng dịch vụ.

Ngoài ra, nghiên cứu còn tập trung vào phân tích doanh thu và tiền tip của khách hàng theo nhiều yếu tố khác nhau như thời gian, khoảng cách di chuyển và phương thức thanh toán. Điều này giúp đánh giá hành vi chi tiêu của khách hàng, cũng như tác động của từng yếu tố đến mức tiền tip mà tài xế nhận được.

Cuối cùng, dự án áp dụng mô hình hóa để dự đoán giá cước taxi dựa trên quãng đường di chuyển. Việc xây dựng mô hình dự đoán có thể giúp khách hàng ước tính chi phí chuyến đi trước khi đặt xe, đồng thời hỗ trợ các công ty taxi điều chỉnh giá phù hợp với tình hình thực tế, tạo ra một hệ thống giá minh bạch và hợp lý hơn.

Để đảm bảo khả năng xử lý dữ liệu lớn một cách hiệu quả, nghiên cứu này ứng dụng các công nghệ và thư viện mạnh mẽ trong lĩnh vực phân tích dữ liệu và học máy.

Trước tiên, **Apache Spark (PySpark)** được sử dụng làm công cụ xử lý dữ liệu lớn. Với khả năng tính toán phân tán, Spark giúp tăng tốc độ xử lý dữ liệu taxi có quy mô hàng triệu bản ghi, đảm bảo hiệu suất cao hơn so với các phương pháp truyền thống. PySpark, giao diện Python của Apache Spark, cung cấp các công cụ cần thiết để làm sạch, tiền xử lý và phân tích dữ liệu một cách linh hoạt.

Tiếp theo, để trực quan hóa dữ liệu, nghiên cứu sử dụng **Seaborn** và **Matplotlib** – hai thư viện mạnh mẽ giúp tạo ra các biểu đồ trực quan, hỗ trợ việc khám phá và hiểu sâu hơn về các xu hướng, mối quan hệ trong dữ liệu. Các biểu đồ như barplot, lineplot, heatmap, scatterplot sẽ được sử dụng để minh họa xu hướng di chuyển của taxi, sự phân bố doanh thu và tiền tip theo nhiều yếu tố khác nhau.

Cuối cùng, để xây dựng mô hình dự đoán giá cước, nghiên cứu áp dụng **Scikit-learn**, một thư viện phổ biến trong lĩnh vực học máy. Scikit-learn cung cấp các thuật toán như hồi quy tuyến tính (Linear Regression) và cây quyết định (Decision Tree), giúp dự báo giá cước taxi dựa trên các đặc trưng như quãng đường, thời gian di chuyển và các yếu tố khác. Việc áp dụng mô hình dự đoán sẽ hỗ trợ khách hàng và nhà cung cấp dịch vụ taxi trong việc đưa ra các quyết định tài chính chính xác hơn.

**CHƯƠNG 2. MÔ TẢ TẬP DỮ LIỆU VÀ CÔNG NGHỆ SỬ DỤNG**

***2.1.* Công nghệ sử dụng**

Để đảm bảo khả năng xử lý dữ liệu lớn một cách hiệu quả, nghiên cứu này ứng dụng các công nghệ và thư viện mạnh mẽ trong lĩnh vực phân tích dữ liệu và học máy.

Trước tiên, **Apache Spark (PySpark)** được sử dụng làm công cụ xử lý dữ liệu lớn. Với khả năng tính toán phân tán, Spark giúp tăng tốc độ xử lý dữ liệu taxi có quy mô hàng triệu bản ghi, đảm bảo hiệu suất cao hơn so với các phương pháp truyền thống. PySpark, giao diện Python của Apache Spark, cung cấp các công cụ cần thiết để làm sạch, tiền xử lý và phân tích dữ liệu một cách linh hoạt.

Tiếp theo, để trực quan hóa dữ liệu, nghiên cứu sử dụng **Seaborn** và **Matplotlib** – hai thư viện mạnh mẽ giúp tạo ra các biểu đồ trực quan, hỗ trợ việc khám phá và hiểu sâu hơn về các xu hướng, mối quan hệ trong dữ liệu. Các biểu đồ như barplot, lineplot, heatmap, scatterplot sẽ được sử dụng để minh họa xu hướng di chuyển của taxi, sự phân bố doanh thu và tiền tip theo nhiều yếu tố khác nhau.

Cuối cùng, để xây dựng mô hình dự đoán giá cước, nghiên cứu áp dụng **Scikit-learn**, một thư viện phổ biến trong lĩnh vực học máy. Scikit-learn cung cấp các thuật toán như hồi quy tuyến tính (Linear Regression) và cây quyết định (Decision Tree), giúp dự báo giá cước taxi dựa trên các đặc trưng như quãng đường, thời gian di chuyển và các yếu tố khác. Việc áp dụng mô hình dự đoán sẽ hỗ trợ khách hàng và nhà cung cấp dịch vụ taxi trong việc đưa ra các quyết định tài chính chính xác hơn.

***2.2* . Dữ liệu và tiền xử lý**

***2.2.1* Dữ liệu sử dụng**

Dữ liệu được sử dụng trong nghiên cứu này đến từ **New York City Taxi & Limousine Commission (TLC)**, một trong những nguồn dữ liệu mở lớn nhất về dịch vụ taxi tại thành phố New York. Tập dữ liệu chứa thông tin chi tiết về hàng triệu chuyến taxi, được thu thập trong một khoảng thời gian dài, giúp cung cấp cái nhìn tổng quan về hoạt động di chuyển của taxi trong thành phố.

Các thông tin quan trọng trong tập dữ liệu bao gồm:

* *Thời gian đón và trả khách*: Ghi nhận thời gian chính xác khi hành khách lên và xuống xe, giúp phân tích xu hướng di chuyển theo thời gian trong ngày và ngày trong tuần.
* *Khoảng cách chuyến đi (miles)*: Thể hiện quãng đường di chuyển của mỗi chuyến, giúp xác định mức độ di chuyển giữa các khu vực khác nhau trong thành phố.
* *Giá cước và tiền tip*: Bao gồm giá cước cơ bản, các khoản phụ phí và tiền tip của khách hàng, giúp nghiên cứu hành vi thanh toán và mức độ hài lòng của khách.
* *Phí cầu đường (tolls) và các khoản phí bổ sung*: Gồm phí tắc nghẽn, phí sân bay và thuế, giúp đánh giá ảnh hưởng của các yếu tố này đến tổng chi phí chuyến đi.
* *Phương thức thanh toán*: Gồm tiền mặt, thẻ tín dụng và các phương thức khác, giúp phân tích xu hướng thanh toán của khách hàng.

Dữ liệu này đóng vai trò quan trọng trong việc phân tích xu hướng di chuyển của taxi, đánh giá doanh thu, và xây dựng mô hình dự đoán giá cước.

***2.2.2. T* Tiền xử lý dữ liệu**

Trước khi tiến hành phân tích và mô hình hóa, dữ liệu cần được tiền xử lý nhằm đảm bảo tính chính xác, nhất quán và phù hợp với các phương pháp phân tích. Các bước tiền xử lý bao gồm:

**Loại bỏ dữ liệu không hợp lệ**

* *Chuyến đi có quãng đường bằng 0:* Những chuyến đi có quãng đường bằng 0 dặm có thể là lỗi nhập liệu hoặc giao dịch không thành công, nên cần được loại bỏ.
* *Giá cước hoặc tiền tip âm:* Giá trị âm không hợp lệ đối với các giao dịch taxi, có thể do lỗi ghi nhận dữ liệu hoặc lỗi hệ thống, cần được loại bỏ để đảm bảo tính chính xác của phân tích.

**Chuyển đổi định dạng thời gian**

* Chuyển đổi các cột thời gian như thời gian đón khách (pickup\_datetime) và thời gian trả khách (dropoff\_datetime) sang định dạng datetime để dễ dàng thao tác.
* Trích xuất các thông tin quan trọng như giờ trong ngày, ngày trong tuần, tháng để phục vụ cho phân tích xu hướng di chuyển theo thời gian.

**Chuẩn hóa dữ liệu**

* *Loại bỏ giá trị ngoại lệ:* 
  + Sử dụng phương pháp IQR (Interquartile Range) để xác định và loại bỏ các giá trị ngoại lệ trong giá cước và tiền tip.
  + Chỉ giữ lại các chuyến đi có giá cước và tiền tip nằm trong khoảng hợp lý.
* *Chuyển đổi dữ liệu dạng phân loại sang số:* 
  + Nếu có các biến phân loại như phương thức thanh toán, chúng cần được chuyển thành dạng số để sử dụng trong mô hình dự đoán.
  + Ví dụ: Tiền mặt (Cash) → 0, Thẻ tín dụng (Credit Card) → 1, Khác → 2.

Sau khi hoàn tất quá trình tiền xử lý, dữ liệu sẽ sạch và có cấu trúc phù hợp để tiến hành các phân tích và mô hình hóa tiếp theo.

# **CHƯƠNG 3. KẾT QUẢ XỬ LÝ, PHÂN TÍCH DỮ LIỆU VÀ ỨNG DỤNG**

***3.1.*Phân tích dữ liệu**

Sau khi tiền xử lý dữ liệu, quá trình phân tích tập trung vào việc tìm hiểu xu hướng di chuyển của taxi tại New York. Các biểu đồ trực quan hóa giúp làm rõ các mô hình di chuyển phổ biến theo thời gian và địa điểm

***3.1.1* Phân tích xu hướng di chuyển**

**Xu hướng số chuyến đi theo giờ trong ngày**

**A graph of blue and white bars

Description automatically generated with medium confidence**

Số lượng chuyến đi có sự biến động rõ rệt theo thời gian trong ngày. Buổi sáng từ 7h đến 9h và buổi tối từ 17h đến 20h là khoảng thời gian cao điểm với lượng khách sử dụng taxi nhiều nhất. Đây là những khung giờ mà người dân di chuyển đến nơi làm việc hoặc trở về nhà, phản ánh đặc trưng giao thông đô thị. Ngược lại, vào khoảng 3h đến 5h sáng, số lượng chuyến đi giảm mạnh do nhu cầu đi lại vào thời điểm này khá thấp.

**So sánh số chuyến giữa ngày thường và cuối tuần**

**A graph of green bars

Description automatically generated**

Hoạt động taxi vào các ngày trong tuần diễn ra khá ổn định, không có sự chênh lệch lớn giữa các ngày. Tuy nhiên, vào cuối tuần, đặc biệt là tối thứ Bảy, số lượng chuyến đi có xu hướng tăng lên. Điều này có thể liên quan đến nhu cầu di chuyển phục vụ mục đích vui chơi, giải trí. Trong khi đó, các ngày đầu tuần lại có mức độ sử dụng taxi thấp hơn, phản ánh một chu kỳ di chuyển quen thuộc của cư dân thành phố.

**Mật độ chuyến đi theo giờ - ngày trong tuần**

**A screen shot of a chart

Description automatically generated**

Sự phân bố chuyến đi theo thời gian được thể hiện rõ qua biểu đồ mật độ. Các giờ cao điểm trong tuần thường có mật độ dày đặc, đặc biệt vào buổi sáng và chiều tối. Trong khi đó, vào cuối tuần, lượng chuyến đi trải rộng hơn trong ngày, không tập trung vào các khung giờ nhất định như ngày thường. Điều này cho thấy sự khác biệt trong thói quen di chuyển giữa ngày làm việc và ngày nghỉ. Những khoảng thời gian ít hoạt động nhất thường rơi vào rạng sáng, khi nhu cầu đi lại giảm xuống mức tối thiểu.

Tổng quan, xu hướng di chuyển taxi tại New York thể hiện rõ tác động của lịch trình làm việc và nhu cầu xã hội. Việc nắm bắt các mô hình này không chỉ giúp tối ưu hóa hoạt động taxi mà còn cung cấp cơ sở cho việc điều chỉnh chính sách giao thông trong thành phố.

***3.1.2* Phân tích điểm đón phổ biến**

**Top 10 địa điểm đón khách phổ biến nhất**

**A graph of different colored bars

Description automatically generated**

Dữ liệu cho thấy một số khu vực có tần suất đón khách cao hơn hẳn so với các địa điểm khác, phản ánh nhu cầu di chuyển mạnh mẽ tại những điểm này. Các địa điểm đón khách phổ biến nhất thường tập trung ở các trung tâm thương mại, khu vực văn phòng và các điểm giao thông lớn như nhà ga và sân bay. Đây là những nơi có lưu lượng người qua lại cao, dẫn đến nhu cầu sử dụng taxi liên tục trong ngày.

Các khu vực trung tâm Manhattan, bao gồm Times Square, Midtown và Financial District, luôn đứng đầu danh sách những điểm đón khách nhiều nhất. Sân bay JFK và LaGuardia cũng góp mặt trong danh sách này, cho thấy vai trò quan trọng của taxi trong việc phục vụ nhu cầu di chuyển từ và đến sân bay. Điều này không chỉ phản ánh tần suất sử dụng dịch vụ cao tại các địa điểm này mà còn cho thấy taxi là một phần quan trọng trong chuỗi vận tải đô thị, kết nối giữa các khu vực trong thành phố.

**So sánh nhu cầu đón khách giữa các quận của NYC**

Bức tranh tổng thể về nhu cầu đón khách giữa các quận tại New York cho thấy sự khác biệt đáng kể giữa các khu vực. Manhattan vẫn là trung tâm chính với số lượng chuyến đi cao nhất, do đây là khu vực tập trung nhiều hoạt động kinh tế, thương mại và du lịch. Các tuyến phố đông đúc tại Midtown và Lower Manhattan thường xuyên có lượng taxi hoạt động lớn nhất.

Trong khi đó, Brooklyn và Queens có số lượng đón khách thấp hơn so với Manhattan nhưng vẫn duy trì mức độ ổn định, đặc biệt là gần các trạm tàu điện ngầm và những khu vực dân cư đông đúc. Khu vực The Bronx và Staten Island có nhu cầu taxi thấp nhất, có thể do hệ thống giao thông công cộng tại đây đã đáp ứng phần lớn nhu cầu đi lại của người dân hoặc do mật độ dân số và hoạt động kinh tế ở những khu vực này thấp hơn.

Tóm lại, phân tích này cho thấy rằng nhu cầu taxi tại New York không phân bổ đồng đều giữa các quận mà phụ thuộc nhiều vào đặc điểm kinh tế, dân cư và hệ thống giao thông tại từng khu vực. Điều này có thể là cơ sở để tối ưu hóa hệ thống phân phối xe, đảm bảo khả năng phục vụ khách hàng tốt hơn tại những khu vực có nhu cầu cao

***3.1.3* Phân tích doanh thu và tip**

**Tổng doanh thu theo giờ trong ngày**

Phân tích doanh thu theo giờ trong ngày cho thấy sự chênh lệch rõ rệt giữa các khung giờ. Nhìn chung, doanh thu taxi tăng cao vào giờ cao điểm buổi sáng (từ 7h đến 9h) và buổi tối (từ 17h đến 20h), khi người dân di chuyển đến nơi làm việc hoặc về nhà. Các khoảng thời gian này cũng trùng với giờ cao điểm giao thông, góp phần làm tăng thời gian di chuyển và kéo theo sự gia tăng doanh thu.

A graph with purple lines

Description automatically generated

Bên cạnh đó, vào ban đêm từ 22h đến 2h sáng, doanh thu cũng có sự tăng vọt đáng kể. Điều này có thể được giải thích bởi nhu cầu di chuyển của khách hàng từ các quán bar, nhà hàng hoặc các sự kiện giải trí. Trái lại, vào khoảng 3h đến 5h sáng, doanh thu chạm mức thấp nhất trong ngày do lượng khách giảm mạnh.

**Mức tip trung bình theo giờ**

Phân tích tiền tip cho thấy mức độ hào phóng của khách hàng thay đổi theo thời gian trong ngày. Thời điểm mà khách hàng tip cao nhất thường rơi vào buổi tối và khuya, đặc biệt trong khoảng 20h đến 1h sáng. Điều này có thể liên quan đến việc khách hàng trong khung giờ này chủ yếu đi chơi, dự tiệc hoặc sử dụng dịch vụ cao cấp, dẫn đến xu hướng tip nhiều hơn.

A graph with red lines

Description automatically generated

Ngược lại, trong giờ cao điểm sáng và trưa, mức tip có xu hướng thấp hơn. Một giả thuyết hợp lý là khách hàng trong khung giờ này thường di chuyển với mục đích công việc, sử dụng taxi như một phương tiện cần thiết thay vì một dịch vụ tiện ích, do đó ít có xu hướng tip cao.

**Quan hệ giữa quãng đường và giá cước trung bình**

Khi phân tích mối quan hệ giữa quãng đường đi và giá cước trung bình thông qua biểu đồ **Scatterplot**, có thể thấy một xu hướng tuyến tính rõ ràng: giá cước tăng theo khoảng cách di chuyển. Tuy nhiên, có một số điểm ngoại lệ với mức giá cước cao bất thường so với quãng đường đi, có thể do các yếu tố như phụ phí sân bay, tắc đường kéo dài hoặc các tuyến đường đặc biệt có phí cầu đường cao.

A group of blue dots

Description automatically generated

Một số chuyến đi ngắn cũng có giá cước cao hơn mức trung bình, có thể do tính năng giá tối thiểu hoặc phí bổ sung tại các địa điểm đón khách quan trọng như sân bay hoặc trung tâm thành phố. Điều này cho thấy rằng không chỉ quãng đường, mà nhiều yếu tố khác cũng ảnh hưởng đến giá cước taxi tại New York, điều mà mô hình dự đoán giá cước sẽ giúp làm rõ hơn trong phần tiếp theo.

***3.1.4* Trực quan hóa bổ sung**

**Phân phối giá cước taxi**

**A graph with a blue line

Description automatically generated**

Phân tích phân phối giá cước taxi thông qua biểu đồ **Histogram** cho thấy phần lớn các chuyến đi có giá cước dao động trong khoảng từ **5 đến 25 USD**. Đây là mức giá phổ biến cho các chuyến đi nội đô hoặc di chuyển ngắn giữa các quận.

Một số ít chuyến có giá cước rất cao, thậm chí lên tới **100 USD hoặc hơn**, tạo ra một đuôi dài phía bên phải của biểu đồ. Những chuyến này có thể là các hành trình dài hơn, chẳng hạn như từ trung tâm thành phố đến các sân bay quốc tế (JFK hoặc Newark) hoặc do có thêm phụ phí như tắc đường và cầu đường.

**Phát hiện ngoại lệ trong giá cước**

Biểu đồ **Boxplot** giúp xác định các giá trị ngoại lệ trong dữ liệu giá cước. Kết quả cho thấy có một số lượng nhỏ các chuyến đi có giá cước **cao đột biến**, vượt xa mức trung bình. Các ngoại lệ này có thể xuất phát từ những chuyến đi đặc biệt như:

* *Chuyến đi đường dài:* Ví dụ từ Manhattan đến các khu vực xa như Long Island hoặc New Jersey.
* *Phụ phí đặc biệt:* Sân bay, tắc đường kéo dài, hoặc dịch vụ cao cấp như xe limousine.
* *Lỗi dữ liệu:* Một số giá trị có thể bị nhập sai hoặc tính toán không chính xác trong quá trình ghi nhận dữ liệu.

**A graph with numbers and lines

Description automatically generated**

Những điểm ngoại lệ này sẽ cần được xem xét kỹ lưỡng trong quá trình mô hình hóa để tránh ảnh hưởng đến độ chính xác của dự đoán giá cước.

**Tỷ lệ chuyến đi được chia sẻ so với chuyến đi đơn lẻ**

Biểu đồ **Pie Chart** thể hiện tỷ lệ chuyến đi taxi được chia sẻ (**Shared Rides**) so với chuyến đi đơn lẻ (**Private Rides**). Kết quả cho thấy phần lớn các chuyến đi vẫn là chuyến đi riêng, trong khi số lượng chuyến đi chia sẻ chiếm một tỷ lệ nhỏ hơn.

A blue circle with a yellow triangle

Description automatically generated

Dịch vụ chia sẻ chuyến đi giúp giảm chi phí cho hành khách nhưng không phổ biến bằng chuyến đi đơn lẻ, có thể do:

* *Tâm lý khách hàng:* Người dân New York thường ưu tiên sự tiện lợi và thời gian hơn là tiết kiệm chi phí.
* *Lịch trình không đồng bộ:* Khó có thể tìm được khách hàng khác có hành trình tương tự cùng thời điểm.
* *Chính sách giá cước:* Sự khác biệt về giá giữa chuyến đi chia sẻ và chuyến đi riêng không quá lớn để đủ hấp dẫn khách hàng lựa chọn dịch vụ chia sẻ.

**Phân phối tip theo giờ trong ngày**

Biểu đồ **Violin Plot** cho thấy sự phân phối tiền tip theo từng giờ trong ngày, giúp làm rõ không chỉ giá trị trung bình mà còn cả sự biến thiên trong hành vi tip của khách hàng.

A graph of a graph

Description automatically generated with medium confidence

Phân tích cho thấy:

* Buổi tối và đêm muộn (20h - 2h sáng) có mức tip cao nhất và biến động lớn nhất. Điều này có thể liên quan đến khách hàng đi nhà hàng, quán bar hoặc dịch vụ cao cấp.
* Giờ cao điểm sáng và trưa (7h - 12h) có mức tip thấp hơn, cho thấy khách hàng đi làm hoặc công tác ít có xu hướng tip cao.
* Một số giờ có giá trị ngoại lệ rất cao, có thể do khách hàng đặc biệt hào phóng hoặc lỗi dữ liệu ghi nhận.

Những phát hiện từ trực quan hóa bổ sung giúp làm sáng tỏ xu hướng chi tiêu của khách hàng, hỗ trợ cho việc mô hình hóa giá cước và dự đoán mức tip trong các phần sau.

### *3.2.* Mô hình hóa - Dự đoán giá cước taxi

***3.2.1* Mục tiêu**

Dự án nhằm xây dựng một mô hình dự đoán giá cước taxi dựa trên các đặc điểm quan trọng như **quãng đường di chuyển, thời gian đón khách, và ngày trong tuần**. Ngoài ra, mô hình cũng được sử dụng để **dự đoán tiền tip của khách hàng** dựa trên các yếu tố tương tự. Việc đánh giá hiệu suất mô hình sẽ giúp kiểm tra độ chính xác và khả năng tổng quát hóa của mô hình đối với dữ liệu thực tế.

***3.2.2* Tiền xử lý dữ liệu cho mô hình**

Dữ liệu thô từ hệ thống taxi chứa nhiều thông tin, nhưng không phải tất cả đều hữu ích cho mô hình dự đoán. Do đó, việc tiền xử lý là bước quan trọng nhằm loại bỏ các dữ liệu không hợp lệ và trích xuất các đặc trưng quan trọng. Đầu tiên, dữ liệu được lọc để loại bỏ các chuyến đi có quãng đường bằng 0, các giao dịch có giá cước hoặc tiền tip mang giá trị âm, vì những trường hợp này có thể xuất phát từ lỗi nhập dữ liệu hoặc sự cố trong hệ thống.

Sau khi dữ liệu được làm sạch, thông tin thời gian từ thời điểm đón khách (pickup\_datetime) được chuyển đổi thành các đặc trưng quan trọng như giờ trong ngày, ngày trong tuần và tháng. Việc này giúp mô hình nhận diện được xu hướng biến động của giá cước theo từng thời điểm, chẳng hạn như vào giờ cao điểm hay ngày cuối tuần, giá taxi có thể khác biệt đáng kể so với những thời điểm khác.

Tiếp theo, các biến quan trọng như quãng đường di chuyển (trip\_miles), thời gian đón khách (pickup\_hour, pickup\_day), giá cước (base\_passenger\_fare) và tiền tip (tips) được lựa chọn để đưa vào mô hình. Những biến này đóng vai trò then chốt trong việc dự đoán giá cước, vì chúng phản ánh trực tiếp yếu tố ảnh hưởng đến chi phí di chuyển.

Trước khi huấn luyện mô hình, các đặc trưng đầu vào cần được chuẩn bị thông qua bước chuyển đổi dữ liệu. Tất cả các biến số được kết hợp thành một vector đầu vào bằng **VectorAssembler**, giúp mô hình dễ dàng tiếp nhận dữ liệu dưới dạng phù hợp. Sau đó, dữ liệu tiếp tục được chuẩn hóa bằng **StandardScaler** nhằm đưa các giá trị về cùng một đơn vị đo lường, tránh tình trạng một biến có giá trị lớn lấn át các biến khác, giúp mô hình học tốt hơn và đưa ra dự đoán chính xác hơn.

***3.2.3* Huấn luyện mô hình**

Mô hình hồi quy tuyến tính được lựa chọn để dự đoán giá cước taxi dựa trên các đặc trưng quan trọng đã qua tiền xử lý. Trước khi huấn luyện, dữ liệu được chia thành hai phần: **80% dành cho huấn luyện** và **20% dành cho kiểm tra**. Việc này giúp mô hình học được các quy luật từ dữ liệu huấn luyện và đánh giá độ chính xác trên tập kiểm tra chưa từng thấy trước đó.

Với bài toán dự đoán giá cước, mô hình **Linear Regression** được huấn luyện bằng cách sử dụng các đặc trưng đầu vào đã chuẩn hóa, bao gồm quãng đường di chuyển, thời gian đón khách theo giờ và ngày trong tuần. Mô hình học cách dự đoán giá cước thực tế dựa trên những yếu tố này, sau đó được kiểm tra bằng cách so sánh kết quả dự đoán với giá trị thực tế từ tập kiểm tra.

Ngoài ra, một mô hình hồi quy tuyến tính khác được sử dụng để dự đoán tiền tip của khách hàng. Quy trình huấn luyện tương tự như với giá cước, nhưng lần này đầu ra của mô hình là biến **tips** – biểu thị số tiền mà khách hàng để lại cho tài xế. Sau khi mô hình được huấn luyện, nó được kiểm tra trên tập dữ liệu kiểm tra để đánh giá mức độ chính xác trong việc dự đoán thói quen tip của hành khách.

Cả hai mô hình đều được đánh giá để kiểm tra xem chúng có thể đưa ra dự đoán sát với giá trị thực tế hay không, qua đó giúp phân tích sâu hơn về yếu tố ảnh hưởng đến giá cước và tiền tip trong hệ thống taxi NYC.

***3.2.4* Đánh giá mô hình**

Để kiểm tra hiệu suất dự đoán của mô hình, chỉ số **RMSE (Root Mean Squared Error)** được sử dụng nhằm đo lường độ chênh lệch trung bình giữa giá trị thực tế và giá trị dự đoán. RMSE giúp đánh giá mức độ sai số trong dự báo, với giá trị càng nhỏ đồng nghĩa với mô hình có độ chính xác cao hơn.

🔹 RMSE (Root Mean Squared Error): 15.224935557487457

+-----------------+-------------------+--------------------+

| prediction|base\_passenger\_fare| scaled\_features|

+-----------------+-------------------+--------------------+

|8.996823869726718| 7.37|[0.00116075259446...|

| 7.37651539438337| 4.73|[0.00165821799209...|

| 6.73956812915588| 8.02|[0.00165821799209...|

|7.164894588641845| 7.19|[0.00165821799209...|

|8.227168313355298| 9.78|[0.00165821799209...|

|8.369291274517773| 6.08|[0.00165821799209...|

|8.511414235680245| 9.77|[0.00165821799209...|

|9.290484462070209| 7.19|[0.00165821799209...|

|9.362067154652175| 16.32|[0.00165821799209...|

|9.291526886071667| 10.91|[0.00165821799209...|

+-----------------+-------------------+--------------------+

only showing top 10 rows

🔹 RMSE cho Tip Prediction: 36.568670671421096

+--------------------+----+--------------------+

| prediction|tips| scaled\_features|

+--------------------+----+--------------------+

| 0.1762426340029094| 0.0|[0.00116075259446...|

| 0.00928124632958767| 0.0|[0.00165821799209...|

|-0.02045988644454...| 0.0|[0.00165821799209...|

|0.020494068792476977| 0.0|[0.00165821799209...|

| 0.09118915680363442| 0.0|[0.00165821799209...|

| 0.11540374190977586| 0.0|[0.00165821799209...|

| 0.1396183270159173| 0.0|[0.00165821799209...|

| 0.19357404489619284| 0.0|[0.00165821799209...|

| 0.22152623748996403| 0.0|[0.00165821799209...|

| 0.22526384497759383| 0.0|[0.00165821799209...|

+--------------------+----+--------------------+

only showing top 10 rows

Kết quả đánh giá cho thấy mô hình dự đoán giá cước taxi đạt **RMSE = 15.22**, cho thấy mức độ sai số tương đối ổn định. Tuy nhiên, mô hình dự đoán tiền tip có **RMSE = 36.57**, phản ánh mức độ biến động lớn của dữ liệu tiền tip. Điều này có thể xuất phát từ việc tip của khách hàng không chỉ phụ thuộc vào quãng đường hay thời gian di chuyển mà còn bị ảnh hưởng bởi các yếu tố chủ quan như thái độ tài xế, trải nghiệm chuyến đi hoặc tình trạng giao thông.

Nhìn chung, mô hình dự đoán giá cước hoạt động khá tốt trong việc ước lượng chi phí chuyến đi, trong khi dự đoán tiền tip vẫn gặp nhiều khó khăn do tính chất không ổn định của dữ liệu thực tế.

***3.2.5* Kết quả dự đoán**

Sau khi huấn luyện mô hình, kết quả dự đoán được so sánh với dữ liệu thực tế để đánh giá mức độ chính xác. Bảng dưới đây thể hiện một số kết quả mẫu của mô hình dự đoán giá cước, bao gồm giá trị thực tế, giá trị dự đoán và các đặc trưng đầu vào.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Giá cước thực tế ($)** | **Giá cước dự đoán ($)** | **Quãng đường (miles)** | **Giờ đón** | **Ngày trong tuần** |
| 7.37 | 8.99 | 0.5 | 14 | 3 |
| 4.73 | 7.37 | 0.4 | 9 | 5 |
| 8.02 | 6.74 | 1.2 | 19 | 6 |
| 7.19 | 7.16 | 0.8 | 12 | 2 |
| 9.78 | 8.23 | 1.5 | 17 | 4 |

Nhìn chung, mô hình dự đoán giá cước khá sát với giá trị thực tế trong nhiều trường hợp, đặc biệt với các chuyến đi có quãng đường ngắn. Tuy nhiên, vẫn có một số sai số đáng kể, đặc biệt với những chuyến đi có giá cước cao hơn mức trung bình.

Với dự đoán tiền tip, sai số lớn hơn đáng kể. Do tiền tip không chỉ bị ảnh hưởng bởi các yếu tố như quãng đường hay thời gian mà còn phụ thuộc vào hành vi của khách hàng, nên mô hình hồi quy tuyến tính gặp khó khăn trong việc ước lượng chính xác. Dữ liệu thực tế cho thấy nhiều chuyến đi có tip bằng 0, trong khi mô hình vẫn dự đoán giá trị dương, làm tăng sai số. Điều này giải thích tại sao RMSE của mô hình dự đoán tiền tip lại cao hơn nhiều so với dự đoán giá cước.

Tóm lại, mô hình hồi quy tuyến tính hoạt động khá tốt trong dự đoán giá cước taxi nhưng gặp nhiều hạn chế trong dự đoán tiền tip do tính biến động cao của dữ liệu thực tế.

**KẾT LUẬN**

***4*. Kết luận và đề xuất**

***4.1* Kết quả đạt được**

Quá trình phân tích dữ liệu taxi NYC đã giúp làm sáng tỏ nhiều khía cạnh quan trọng trong hành vi di chuyển của hành khách. Thông qua việc trực quan hóa dữ liệu, có thể nhận thấy rõ các xu hướng di chuyển theo thời gian, bao gồm sự biến động số lượng chuyến đi theo từng khung giờ trong ngày và sự khác biệt giữa các ngày trong tuần. Các phân tích cũng chỉ ra rằng vào giờ cao điểm buổi sáng và chiều tối, nhu cầu sử dụng taxi tăng cao đáng kể, phản ánh thói quen di chuyển của cư dân thành phố.

Ngoài ra, việc xác định các điểm đón khách phổ biến giúp làm rõ những khu vực có nhu cầu cao đối với dịch vụ taxi. Những khu vực trung tâm thương mại, nhà ga và sân bay thường xuyên có số lượng chuyến đón lớn, cho thấy sự tập trung di chuyển đáng kể tại những địa điểm này. Điều này không chỉ hữu ích cho các công ty taxi trong việc tối ưu hóa phân bổ xe mà còn cung cấp thông tin quan trọng cho việc quy hoạch giao thông đô thị.

Mô hình dự đoán giá cước taxi đã đạt được độ chính xác tương đối cao khi đánh giá bằng RMSE. Kết quả cho thấy quãng đường đi là yếu tố quan trọng nhất quyết định giá cước, trong khi yếu tố thời gian trong ngày cũng ảnh hưởng đến sự chênh lệch về mức giá. Mặc dù vậy, mô hình gặp khó khăn trong việc dự đoán tiền tip do yếu tố này phụ thuộc nhiều vào tâm lý khách hàng, thói quen chi tiêu và các yếu tố ngẫu nhiên khác.

Nhìn chung, nghiên cứu này đã giúp làm sáng tỏ cách thức hoạt động của dịch vụ taxi NYC thông qua dữ liệu lớn, mang lại những giá trị thực tiễn quan trọng cho cả doanh nghiệp vận tải và các cơ quan quản lý giao thông.

### *4.2.* Hạn chế và hướng phát triển

Mặc dù phân tích dữ liệu taxi NYC đã mang lại nhiều kết quả hữu ích, vẫn tồn tại một số hạn chế cần được khắc phục để nâng cao tính chính xác và ứng dụng thực tiễn của nghiên cứu. Trước tiên, dữ liệu có thể chứa sai sót do các ngoại lệ hoặc lỗi nhập liệu, chẳng hạn như giá trị quãng đường bằng 0 hoặc tiền tip âm. Dù đã áp dụng các bước tiền xử lý để loại bỏ dữ liệu không hợp lệ, vẫn có khả năng một số dữ liệu nhiễu chưa được nhận diện đầy đủ, ảnh hưởng đến kết quả phân tích và dự đoán.

Bên cạnh đó, mô hình dự đoán giá cước chủ yếu dựa trên các yếu tố như quãng đường và thời gian, trong khi thực tế giá taxi có thể bị ảnh hưởng bởi nhiều yếu tố khác, chẳng hạn như điều kiện thời tiết, tình trạng giao thông hoặc các sự kiện đặc biệt diễn ra trong thành phố. Việc tích hợp thêm dữ liệu về thời tiết, sự kiện lớn hay các tuyến đường tắc nghẽn có thể giúp cải thiện độ chính xác của mô hình dự đoán.

Ngoài ra, nghiên cứu hiện tại chủ yếu tập trung vào phân tích xu hướng di chuyển và dự đoán giá cước. Một hướng phát triển tiềm năng là mở rộng phân tích để tối ưu hóa lộ trình và giá cả cho các dịch vụ taxi. Việc sử dụng dữ liệu lịch sử để đề xuất tuyến đường nhanh nhất hoặc chi phí tối ưu có thể mang lại giá trị lớn cho cả hành khách lẫn nhà cung cấp dịch vụ vận tải. Điều này không chỉ giúp giảm thiểu thời gian di chuyển mà còn có thể tăng hiệu suất hoạt động của hệ thống taxi trong thành phố.

Nhìn chung, nghiên cứu này có thể được mở rộng theo nhiều hướng để tăng cường ứng dụng thực tế, từ việc cải thiện mô hình dự đoán đến tối ưu hóa hệ thống giao thông dựa trên dữ liệu lớn.

**DANH MỤC TÀI LIỆU THAM KHẢO**

1. Apache Software Foundation. (2024). *Apache Spark Documentation*. Retrieved from <https://spark.apache.org/docs/latest/>
2. New York City Taxi & Limousine Commission (TLC). (2024). *NYC Taxi Trip Data*. Retrieved from https://www.nyc.gov/html/tlc/html/about/trip\_record\_data.shtml
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python*. *Journal of Machine Learning Research, 12*, 2825-2830.
4. McKinney, W. (2010). *Data structures for statistical computing in Python*. *Proceedings of the 9th Python in Science Conference*, 51-56.
5. Seaborn Development Team. (2024). *Seaborn: Statistical Data Visualization*. Retrieved from https://seaborn.pydata.org/
6. Pedersen, T. L. (2024). *Matplotlib: Visualization with Python*. Retrieved from https://matplotlib.org/stable/contents.html
7. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer.
8. Breiman, L. (2001). *Random forests*. *Machine Learning, 45*(1), 5-32.
9. Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. *Annals of Statistics, 29*(5), 1189-1232.
10. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.
11. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
12. Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.