HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



PHẠM MINH ĐỨC

ỨNG DỤNG MÔ HÌNH DỰ ĐOÁN CHUỖI THỜI GIAN TỐI ƯU CHI PHÍ VẬN TẢI

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI – 2025

Đề án tốt nghiệp được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS.TS. Nguyễn Mạnh Hùng**
(Ghi rõ học hàm, học vị)

Phản biện 1: . . . . . .

Phản biện 2: . . . . . .

Đề án tốt nghiệp sẽ được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng ..... năm .....

1 Lý do chọn đề tài

Trong bối cảnh đô thị hóa nhanh và sự phát triển mạnh mẽ của các dịch vụ vận tải thông minh, việc dự báo chính xác thời gian di chuyển (ETA – Estimated Time of Arrival) của phương tiện trở thành một yếu tố quan trọng trong tối ưu hóa điều hành giao thông, lập kế hoạch vận tải và nâng cao trải nghiệm người dùng. Các hệ thống như Grab, Uber hay dịch vụ taxi truyền thống đều dựa vào mô hình ETA để ước tính thời gian đến, phân bổ tài nguyên hợp lý và cung cấp thông tin chính xác cho khách hàng. Tuy nhiên, đặc trưng dữ liệu giao thông đô thị luôn biến động, chịu ảnh hưởng của nhiều yếu tố như mật độ đường, điều kiện thời tiết, thời gian trong ngày và hành vi di chuyển của con người, khiến các phương pháp dự báo truyền thống khó đạt hiệu quả cao. Xuất phát từ thực tế đó, đề tài "Dự đoán thời gian di chuyển của phương tiện vận tải hành khách bằng mô hình học máy XGBoost" được lựa chọn nhằm xây dựng và kiểm chứng một mô hình dự báo có độ chính xác cao, khả năng mở rộng và ứng dụng thực tiễn trong các hệ thống giao thông thông minh.

2 Tổng quan về vấn đề nghiên cứu

Nhiều nghiên cứu trong và ngoài nước đã tập trung vào bài toán dự báo ETA dưới nhiều góc độ khác nhau. Các phương pháp truyền thống dựa trên mô hình thống kê như ARIMA, SARIMA hay Kalman Filter có ưu điểm trong việc mô hình hóa chuỗi thời gian nhưng lại bị hạn chế bởi giả định tuyến tính và không thể nắm bắt được các mối quan hệ phi tuyến giữa các biến đầu vào. Với sự phát triển của học máy, các mô hình như Random Forest, Gradient Boosting, XGBoost và LightGBM đã chứng minh khả năng vượt trội trong việc khai thác các đặc trưng phức tạp của dữ liệu giao thông. Đặc biệt, XGBoost nổi bật với cơ chế học tăng cường gradient, khả năng regularization mạnh mẽ và tính mở rộng cao, cho phép xử lý tập dữ liệu lớn như bộ NYC Taxi 2024. Trong khi đó, các mô hình học sâu như LSTM hay Graph Neural Network (GNN) mang lại độ chính xác cao hơn trong một số kịch bản, song đòi hỏi tài nguyên tính toán lớn và thời gian huấn luyện dài, hạn chế khả năng triển khai thực tế. Do đó, XGBoost được lựa chọn làm phương pháp nghiên cứu chính trong đề tài này như một giải pháp cân bằng giữa hiệu năng, độ chính xác và khả năng giải thích kết quả.

3 Mục đích nghiên cứu

Mục đích của đề tài là xây dựng một mô hình học máy có khả năng dự báo chính xác thời gian di chuyển của phương tiện trong môi trường giao thông đô thị phức tạp, thông qua việc khai thác hiệu quả các đặc trưng không gian–thời gian và các yếu tố ngữ cảnh. Cụ thể, nghiên cứu hướng đến việc thiết kế quy trình khai phá dữ liệu có hệ thống, bao gồm làm sạch, xử lý ghép bản đồ, tạo đặc trưng và phân tích tương quan, từ đó huấn luyện mô hình XGBoost để dự báo thời gian di chuyển. Bên cạnh đó, đề tài cũng hướng tới việc đánh giá mô hình thông qua

các chỉ số định lượng như RMSE, MAE và MAPE, đồng thời phân tích ý nghĩa của các đặc trưng bằng phương pháp SHAP nhằm đảm bảo mô hình không chỉ chính xác mà còn có khả năng diễn giải cao, phục vụ tốt cho các ứng dụng thực tiễn trong lĩnh vực vận tải thông minh.

4 Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là các mô hình học máy áp dụng trong dự báo thời gian di chuyển của phương tiện, trong đó tập trung vào mô hình XGBoost và các kỹ thuật khai phá dữ liệu hỗ trợ. Phạm vi nghiên cứu được giới hạn trên bộ dữ liệu taxi công cộng của thành phố New York năm 2024, bao gồm hơn 2 triệu bản ghi về hành trình thực tế. Dữ liệu được xử lý và phân tích nhằm mô hình hóa mối quan hệ giữa các yếu tố như vị trí đón – trả, khoảng cách theo mạng lưới đường, thời gian trong ngày, mật độ giao thông và điều kiện thời tiết với biến mục tiêu là thời lượng chuyến đi. Phạm vi của đồ án tập trung vào việc kiểm chứng tính khả thi của mô hình trong điều kiện dữ liệu lịch sử và bán tĩnh, chưa mở rộng sang dữ liệu thời gian thực hoặc các yếu tố dự báo ngắn hạn như sự kiện bất thường hay thay đổi tuyến đường.

5 Phương pháp nghiên cứu

Đề tài sử dụng phương pháp tiếp cận định lượng kết hợp với kỹ thuật khai phá dữ liệu và học máy hiện đại. Quá trình nghiên cứu bao gồm các bước: (1) thu thập và tiền xử lý dữ liệu từ bộ NYC Taxi Dataset; (2) khai phá dữ liệu để phát hiện và xây dựng các đặc trưng có ý nghĩa, bao gồm đặc trưng không gian, thời gian và ngữ cảnh; (3) huấn luyện mô hình XGBoost với các kỹ thuật tối ưu siêu tham số như Grid Search và Cross-validation; (4) đánh giá mô hình bằng các chỉ số sai số RMSE, MAE, MAPE; và (5) phân tích mô hình bằng công cụ SHAP để xác định vai trò của từng đặc trưng trong dự báo. Cách tiếp cận này cho phép đảm bảo cả độ chính xác, tính ổn định và khả năng diễn giải của mô hình, đồng thời tạo nền tảng vững chắc cho việc mở rộng sang các hướng nghiên cứu trong tương lai như dự báo ETA thời gian thực hoặc mô hình lai giữa học sâu và boosting.

CHƯƠNG 1: BÀI TOÁN DỰ ĐOÁN THỜI GIAN VẬN TẢI HÀNH KHÁCH

1.1 Giới thiệu bài toán dự đoán thời gian vận tải hành khách

1.1.1 Định nghĩa và mục tiêu của bài toán

Bài toán dự đoán thời gian vận tải hành khách (Estimated Time of Arrival – ETA) là thành phần cốt lõi trong hệ thống giao thông thông minh (ITS). Mục tiêu là ước lượng chính xác thời gian phương tiện hoàn thành hành trình dựa trên các đặc trưng không–thời gian như vị trí, thời điểm, điều kiện giao thông. Đề án tập trung vào **dự báo ngắn hạn**, phù hợp cho điều phối thời gian thực trong các dịch vụ taxi, xe buýt và logistics đô thị.

ETA được mô hình hoá như một bài toán hồi quy $f(x) = y$ với x là đặc trưng hành trình và y là thời gian di chuyển. Dự báo chính xác giúp giảm thời gian chờ, tối ưu hóa điều phối và nâng cao hiệu quả khai thác đội xe.

1.1.2 Ứng dụng giải quyết bài toán trong thực tế

- Tối ưu hóa lộ trình và chi phí vận hành.
- Nâng cao trải nghiệm khách hàng nhờ ước lượng thời gian chính xác.
- Hỗ trợ ra quyết định điều phối và phân bổ phương tiện.

1.1.3 Các nghiên cứu liên quan trong lịch sử

Các nghiên cứu ETA phát triển từ mô hình thống kê (ARIMA, Kalman), sang học máy (Random Forest, Gradient Boosting), đến học sâu (LSTM, CNN, Transformer). Các mô hình gần đây khai thác dữ liệu đa nguồn, quỹ đạo GPS và đặc trưng không–thời gian để cải thiện độ chính xác dự báo.

Hai hướng chính trong nghiên cứu: **(1)** mô hình dựa trên đặc trưng (GCN, routing features) kết hợp XGBoost/GBDT; **(2)** mô hình end-to-end như DeepTTE, ST-NN, TETA học trực tiếp từ dữ liệu thô.

Các mô hình Transformer cho thấy ưu thế về khả năng học phụ thuộc dài hạn và tương tác không–thời gian phức tạp [**?**, **?**, **?**, **?**].

1.2 Các phương pháp dự đoán truyền thống trước đây

1.2.1 Phương pháp dựa trên thống kê

a) ARIMA/SARIMA.

ARIMA mô hình hóa giá trị tương lai của chuỗi thời gian thông qua quan sát và sai số quá khứ. SARIMA mở rộng cho dữ liệu có mùa vụ. Mặc dù dễ triển khai và có tính diễn giải

tốt, mô hình tuyến tính này khó mô tả quan hệ phi tuyến và không gian—thời gian trong giao thông đô thị. Công thức tổng quát:

$$\left(1-\sum_{i=1}^{p}\phi_i B^i\right)(1-B)^d Y_t = c+\left(1+\sum_{j=1}^{q}\theta_j B^j\right)\varepsilon_t. \tag{1.1}$$

b) Bộ lọc Kalman.

Kalman Filter ước lượng trạng thái (vị trí, vận tốc) bằng cơ chế dự báo–cập nhật, phù hợp cho dữ liệu GPS nhiễu và ứng dụng thời gian thực. Hạn chế chính là giả định tuyến tính và nhiễu Gaussian, không phù hợp các quan hệ giao thông phi tuyến mạnh.

$$\hat{x}_{t|t-1} = F_t\hat{x}_{t-1|t-1} + B_t u_t, \tag{1.2}$$

$$K_t = P_{t|t-1}H_t^\top (H_t P_{t|t-1}H_t^\top + R_t)^{-1}. \tag{1.3}$$

1.2.2 Phương pháp dựa trên mô hình học máy

a) kNN Regression.

Phương pháp đơn giản, phi tuyến tự nhiên; dự đoán theo trung bình k láng giềng gần nhất:

$$\hat{y}(\mathbf{x}) = \frac{\sum_{i=1}^{k} w_i(\mathbf{x})y_i}{\sum_{i=1}^{k} w_i(\mathbf{x})}. \tag{1.4}$$

Hạn chế: chi phí tính toán cao khi dự báo.

b) Support Vector Regression (SVR).

SVR tối ưu siêu phẳng hồi quy trong biên ε thông qua kernel. Mạnh với dữ liệu vừa nhưng khó tinh chỉnh tham số.

c) CART.

Cây quyết định chia không gian đặc trưng theo ngưỡng tối ưu để giảm SSE. Dễ diễn giải nhưng dễ overfitting; thường làm base learner trong ensemble.

1.3 Các phương pháp dự đoán hiện đại

1.3.1 Ứng dụng mô hình học máy kết hợp (Ensemble Learning)

a) Random Forest.

RF kết hợp nhiều cây CART huấn luyện trên mẫu bootstrap, giảm phương sai và tăng độ ổn định. Dự đoán được tính theo trung bình:

$$\hat{y}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}).$$

b) Gradient Boosting và XGBoost.

Gradient Boosting xây mô hình tuần tự để sửa lỗi vòng trước:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta h_m(\mathbf{x}).$$

XGBoost tối ưu hoá hơn nhờ regularization L1/L2, subsampling, parallelization và là mô hình mạnh trong ETA [**?**].

1.3.2 Ứng dụng mô hình học sâu

a) LSTM.

LSTM xử lý chuỗi thời gian GPS dài, khắc phục vanishing gradient:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \tag{1.5}$$

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t. \tag{1.6}$$

Mạnh với dữ liệu chuỗi nhưng huấn luyện chậm.

b) Transformer và Attention.

Transformer dùng Self-Attention để học quan hệ dài hạn mà không cần tuần tự:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{1.7}$$

Các mô hình DeepETA, TETA đạt kết quả vượt trội trong ETA [**?**, **?**].

1.4 Đề xuất phương hướng

1.4.1 Khai phá dữ liệu trong bài toán ETA

Khai phá dữ liệu đóng vai trò nền tảng trong ETA, giúp xử lý dữ liệu đa nguồn và trích xuất đặc trưng không–thời gian chất lượng cao. Quy trình gồm:

a) Thu thập và tích hợp dữ liệu GPS, traffic, weather.

b) Làm sạch nhiễu GPS, xử lý ngoại lệ.

c) Map–matching bằng HMM để tái tạo quỹ đạo.

d) Tạo đặc trưng không–thời gian, tuyến đường, thời tiết, mật độ.

e) Lựa chọn đặc trưng dựa trên tương quan và SHAP.

1.4.2 Sử dụng học máy kết hợp Ensemble Learning

XGBoost được chọn làm mô hình chính nhờ sự cân bằng giữa độ chính xác, tốc độ và khả năng diễn giải (SHAP). Kết hợp với pipeline đặc trưng chuẩn (khoảng cách, thời gian, đặc trưng OSRM) giúp mô hình đạt hiệu năng cao và ổn định trong dữ liệu taxi đô thị.

1.5 Kết luận chương

Chương 1 đã tổng quan bài toán ETA từ thống kê truyền thống đến học sâu hiện đại. ARIMA/Kalman phù hợp cho dữ liệu tuyến tính nhưng hạn chế trong môi trường giao thông phi tuyến. Các mô hình học sâu (LSTM, Transformer) đạt độ chính xác cao nhưng yêu cầu tài nguyên lớn.

XGBoost – đại diện cho Gradient Boosting – nổi bật nhờ hiệu năng mạnh, khả năng xử lý phi tuyến và mở rộng tốt trên dữ liệu lớn. Cùng với quy trình khai phá dữ liệu chất lượng cao, đây là cơ sở để triển khai mô hình ETA ở Chương 2.

CHƯƠNG 2: HUẤN LUYỆN MÔ HÌNH DỰ BÁO THỜI GIAN VẬN TẢI HÀNH KHÁCH

2.1 Kỹ thuật khai phá dữ liệu

Trên các bộ dữ liệu NYC Taxi (*Trip Duration*, *Taxi Fare*), những lời giải đạt thứ hạng cao trên Kaggle đều nhấn mạnh một quy trình khai phá dữ liệu chặt chẽ, kết hợp tái cấu trúc quỹ đạo, rút gọn hình học, làm giàu ngữ cảnh không–thời gian và bổ sung thông tin mạng lưới đường. Mục tiêu của quy trình là chuyển đổi toạ độ GPS và mốc thời gian thô thành một không gian đặc trưng giàu ý nghĩa, ổn định phương sai và có khả năng khái quát trên các điều kiện giao thông khác nhau.

2.1.1 Xử lý ghép bản đồ – Map matching

Dữ liệu GPS đô thị chịu ảnh hưởng nhiễu đa nguồn, do đó cần ánh xạ các điểm quan sát lên đồ thị đường thực (*road network*) để khử lệch vị trí và khôi phục tuyến đường hợp lý. Cách tiếp cận phổ biến là mô hình ẩn Markov (HMM), trong đó mỗi điểm GPS được quy chiếu sang một tập ứng viên cạnh đường với xác suất phát xạ phụ thuộc sai số đo, còn xác suất chuyển tiếp phản ánh ràng buộc hình học và động học giữa các cạnh kề nhau.

Bên cạnh ánh xạ HMM, các lời giải đứng hạng cao còn khai thác các dịch vụ định tuyến mã nguồn mở (ví dụ OSRM) để suy ra các đại lượng theo mạng lưới, như thời gian lái xe tối ưu, chiều dài lộ trình theo đường, số lần rẽ, số nút giao và số đèn tín hiệu trên tuyến. Những biến này mang tính diễn giải cao hơn khoảng cách đường chim bay, giúp mô hình tiếp cận gần hơn với chi phí thời gian thực tế.

2.1.2 Sử dụng thuật toán Ramer–Douglas–Peucker

Sau ghép bản đồ, quỹ đạo vẫn có thể dày đặc điểm. Thuật toán *Ramer–Douglas–Peucker (RDP)* được áp dụng để nén hình, giữ lại tập điểm "xương sống" đại diện cho hình dạng tuyến với sai số biên kiểm soát được. Việc rút gọn này làm giảm độ phức tạp tính toán cho các phép suy luận theo mạng lưới, đồng thời cải thiện tính ổn định của các đặc trưng hình học (góc phương vị, độ uốn khúc, chỉ số ngoằn ngoèo) trước nhiễu vị trí cục bộ.

2.1.3 Biến đổi dữ liệu và tạo đặc trưng mới

Các đặc trưng không gian cơ bản thường bắt đầu từ khoảng cách hình học giữa điểm đón và trả. Khoảng cách *haversine* mô tả độ dài cung tròn trên mặt cầu, trong khi khoảng cách "Manhattan" (*taxicab distance*) xấp xỉ tốt cho lưới phố trực giao của New York. Góc phương vị (*bearing*) chuẩn hoá hướng chuyển động, còn tỷ số thẳng tuyến (khoảng cách đường thẳng chia cho độ dài lộ trình theo mạng) phản ánh mức độ ngoằn ngoèo.

Trong bối cảnh thành phố có nhiều điểm hấp dẫn (POIs) ảnh hưởng luồng xe, khoảng

cách tới các sân bay lớn (JFK, LGA, EWR), trọng tâm Manhattan hay trung tâm thương mại được sử dụng như biến ngữ cảnh cố định. Các biến nhị phân nhận diện "chuyến sân bay" giúp mô hình học các chế độ lưu thông đặc thù.

Các đặc trưng thời gian đóng vai trò trục chính để mô hình hoá chu kỳ giao thông. Thực hành tốt bao gồm tách giờ trong ngày, ngày trong tuần, tháng trong năm; mã hoá các khung giờ cao điểm sáng/chiều; thêm cờ ngày cuối tuần và ngày nghỉ lễ theo lịch Hoa Kỳ; và, khi sẵn có, kết hợp điều kiện thời tiết đồng bộ theo mốc thời gian (nhiệt độ, lượng mưa/tuyết, gió). Việc áp dụng biến đổi *log1p* cho biến mục tiêu thời gian giúp ổn định phân phối lệch phải và giảm ảnh hưởng ngoại lai, đồng thời chuẩn hoá hoặc chuẩn vị cho các thước đo khoảng cách.

Đặc trưng theo mạng lưới đường cho thấy đóng góp nổi trội khi dự báo thời gian. Bên cạnh thời gian và cự ly lái xe theo OSRM, số lần rẽ trái/phải, số nút giao và số phân đoạn đường (*segments*) trên tuyến là các biến gắn trực tiếp với chi phí trễ thực tế. Tại các nút giao dày đặc, việc bổ sung *proxy* cho mật độ đèn tín hiệu hoặc giới hạn tốc độ theo phân loại đường càng tăng tính mô tả của mô hình.

Đặc trưng mật độ–đếm (*counting features*) được xây dựng bằng cách phân mảnh không gian theo lưới đều, *geohash* hoặc cụm K-means trên toạ độ đón/trả. Với mỗi ô hoặc cụm, ta tính các thống kê lịch sử như số chuyến, tốc độ trung bình, thời gian trung vị theo khung giờ. Các đặc trưng phân cụm còn được tận dụng ở mức cao hơn bằng cách gán nhãn cụm cho điểm đón và trả, rồi xây dựng tương tác "cặp cụm" (*origin cluster* $\times$ *destination cluster*).

Các đặc trưng tương tác được hình thành có chủ đích nhằm nắm bắt phi tuyến giữa không gian và thời gian, tiêu biểu như tích số của khoảng cách với cờ giờ cao điểm, hoặc tương tác giữa cự ly theo mạng và lượng mưa. Các tỷ suất tốc độ theo ô không gian–thời gian cũng giúp ràng buộc dự báo vào bối cảnh giao thông địa phương. Cuối cùng, để giảm rủi ro rò rỉ dữ liệu, mọi đặc trưng dựa trên thống kê lịch sử cần tính trong khung thời gian quá khứ so với bản ghi mục tiêu.

2.1.4 Phân tích mô tả và tính tương quan đặc trưng

Sau khi hình thành không gian đặc trưng, phân tích mô tả giúp thẩm định trực quan phân phối và phát hiện ngoại lai, trong khi phân tích tương quan định lượng đóng vai trò sàng lọc. Hệ số tương quan Pearson hữu ích với quan hệ tuyến tính sau các biến đổi log, còn *mutual information* và độ quan trọng hoán vị (*permutation importance*) phù hợp hơn cho quan hệ phi tuyến.

Trong thực nghiệm với XGBoost, việc diễn giải thông qua *SHAP values* cho thấy các đặc trưng theo mạng lưới (thời gian/độ dài lộ trình OSRM), hướng chuyển động, chỉ báo giờ cao điểm và cụm không gian thường xếp hạng cao về mức đóng góp, trong khi các biến ngữ cảnh như thời tiết và sân bay gia tăng ổn định cho những tuyến có hành vi khác biệt.

2.2 Mô hình dự đoán XGBoost

XGBoost (*Extreme Gradient Boosting*) là một thuật toán học máy mạnh mẽ thuộc họ mô hình *boosting*, được thiết kế nhằm nâng cao hiệu suất dự báo thông qua việc kết hợp nhiều mô hình yếu (weak learners), thường là các cây quyết định nhỏ, thành một mô hình tổng hợp có năng lực khái quát hóa tốt. Trong bài toán dự đoán ETA, XGBoost đặc biệt hiệu quả vì có thể học được các mối quan hệ phi tuyến phức tạp giữa các đặc trưng không–thời gian.

2.2.1 Nguồn gốc mô hình

XGBoost được phát triển nhằm tối ưu hóa tốc độ huấn luyện và khả năng mở rộng trên các tập dữ liệu lớn, đồng thời duy trì độ chính xác cao thông qua cơ chế chuẩn hóa và hỗ trợ tính toán song song. Mô hình này nhanh chóng trở thành chuẩn mực trong các cuộc thi học máy nhờ hiệu năng vượt trội.

XGBoost hỗ trợ nhiều dạng bài toán khác nhau, bao gồm hồi quy, phân loại, xếp hạng và dự báo chuỗi thời gian. Trong lĩnh vực giao thông, mô hình này được sử dụng rộng rãi trong các bài toán ước lượng thời gian hành trình (ETA), dự đoán độ trễ phương tiện công cộng và tối ưu hóa điều phối xe.

2.2.2 Các tính năng chính

XGBoost kế thừa cơ chế học tăng cường dần (*additive boosting*), trong đó các mô hình yếu được huấn luyện liên tiếp để giảm sai số còn lại. Một trong những cải tiến cốt lõi của XGBoost là việc bổ sung điều khoản chuẩn hóa (*L1, L2 regularization*) vào hàm mục tiêu:

$$\mathscr{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.8}$$

với

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j} w_j^2, \tag{2.9}$$

trong đó T là số nút lá và w_j là trọng số.

Ngoài ra, XGBoost còn nổi bật nhờ:

- **Tối ưu hoá phân tách:** thuật toán tìm điểm chia hiệu quả ngay cả khi dữ liệu có trọng số không đều.
- **Xử lý thiếu dữ liệu:** tự động học hướng phân nhánh khi gặp giá trị null.
- **Hỗ trợ tính song song:** tăng tốc quá trình xây dựng cây quyết định.
- **Mở rộng tốt:** xử lý dữ liệu lớn nhờ cấu trúc bộ nhớ tối ưu.

2.2.3 Ứng dụng của mô hình

Trong bài toán dự đoán thời gian di chuyển của phương tiện vận tải, XGBoost đạt hiệu quả cao nhờ khả năng học phi tuyến giữa các đặc trưng như thời gian, vị trí, mật độ giao thông và điều kiện thời tiết.

XGBoost cũng hỗ trợ phân tích diễn giải thông qua *feature importance* và *SHAP values*, giúp mô hình minh bạch hơn và dễ áp dụng trong thực tế.

2.3 Phương pháp đánh giá và thực nghiệm mô hình

2.3.1 Mục tiêu và phạm vi đánh giá

Việc đánh giá mô hình nhằm xác định khả năng dự báo ETA, kiểm định tính ổn định và khả năng tổng quát. Các tiêu chí chính gồm:

- độ chính xác dự báo,
- mức ổn định trên các phân đoạn dữ liệu,
- khả năng diễn giải.

2.3.2 Quy trình huấn luyện và chia tập dữ liệu

Để tối ưu hoá hiệu suất mô hình trên tập dữ liệu quy mô lớn trong điều kiện tài nguyên tính toán giới hạn, bộ dữ liệu được chia thành hai phần: *tập huấn luyện* (Training set) và *tập kiểm định* (Validation set) với tỷ lệ 80% – 20%. Cách phân chia này giúp mô hình có đủ dữ liệu để học các mẫu hình phức tạp, đồng thời vẫn giữ lại một lượng dữ liệu độc lập đủ lớn để đánh giá khả năng tổng quát hoá và hạn chế hiện tượng quá khớp (*overfitting*). Việc sử dụng *random split* là phù hợp với bản chất của bài toán ước lượng thời gian di chuyển, nơi mỗi chuyến đi được xem như một quan sát độc lập thay vì một chuỗi thời gian liên tục.

Biến mục tiêu được xác định là *thời gian di chuyển thực tế* (đơn vị phút). Việc biểu diễn thời gian dưới dạng phút thay vì giây giúp cho các chỉ số đánh giá như RMSE và MAE trở nên trực quan hơn và phản ánh rõ hơn sai số trong thực tế vận hành. Trước khi đưa vào mô hình, dữ liệu được xử lý ngoại lệ và làm giàu đặc trưng (*Feature Engineering*) nhằm chuyển đổi các thuộc tính không gian–thời gian thô thành các biến có ý nghĩa ngữ nghĩa cao, như giờ trong ngày, ngày trong tuần, khoảng cách Haversine, tốc độ trung bình ước tính, và các đặc trưng theo mùa vụ.

Quá trình huấn luyện sử dụng hàm mất mát hồi quy bình phương lỗi (*Squared Error Loss*), được cấu hình thông qua tham số `reg:squarederror` trong thư viện XGBoost. Hàm mất mát này được tối ưu hoá bằng thuật toán *Gradient Boosting*, trong đó mỗi cây quyết định mới được xây dựng nhằm giảm dần phần dư (*residuals*) mà các cây trước đó chưa giải thích được. Cơ chế học tuần tự này giúp mô hình hội tụ hiệu quả và xử lý tốt các đặc trưng phi tuyến trên không gian đặc trưng phức tạp.

2.3.3 Tinh chỉnh siêu tham số

Tinh chỉnh được thực hiện bằng *Grid Search* hoặc *Bayesian Optimization*. Các siêu tham số quan trọng gồm:

- `max_depth`,
- `learning_rate`,
- `n_estimators`,
- `subsample`, `colsample_bytree`,
- `lambda`, `alpha`.

2.3.4 Chỉ số đánh giá hiệu suất mô hình

Các chỉ số đánh giá gồm:

MSE:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

MAE:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|.$$

MAPE: đo sai số theo tỉ lệ phần trăm.

Bên cạnh đó, phân tích SHAP giúp diễn giải đóng góp của từng đặc trưng.

2.3.5 Phương pháp diễn giải và kiểm định mô hình

Mô hình được phân tích bằng *feature importance* và SHAP để hiểu mức độ ảnh hưởng của đặc trưng. Hiệu suất được kiểm định qua nhiều cấu hình siêu tham số khác nhau.

2.4 Kết luận chương

Chương 2 đã trình bày đầy đủ kỹ thuật khai phá dữ liệu, xây dựng đặc trưng, mô hình XGBoost và quy trình đánh giá. Khung phương pháp này là nền tảng cho Chương 3 – nơi các kết quả thực nghiệm sẽ được phân tích chi tiết.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

3.1 Bài toán chi tiết

Chương này mô tả quá trình thực nghiệm mô hình dự báo thời gian di chuyển (ETA) bằng XGBoost. Đây là bài toán hồi quy phi tuyến, đầu vào gồm các đặc trưng không–thời gian và đặc trưng ngữ cảnh, đầu ra là thời lượng chuyến đi theo phút. Mục tiêu là xây dựng mô hình dự báo chính xác, phù hợp dữ liệu giao thông biến động của đô thị.

3.2 Kịch bản tiến hành thực nghiệm

3.2.1 Môi trường thực nghiệm

Thí nghiệm được triển khai trên Databricks Community Edition (CPU: 2 vCPU, RAM 15 GB). Dù không có GPU, Spark hỗ trợ xử lý phân tán hiệu quả. Các thư viện sử dụng gồm PySpark, Pandas, NumPy, Scikit-learn, SHAP và XGBoost. Mô hình XGBoost được huấn luyện bằng XGBoost4J–Spark, còn các mô hình baseline được huấn luyện trên tập mẫu.

3.2.2 Bộ dữ liệu thực nghiệm

Bộ dữ liệu NYC Yellow Taxi 10/2024 gồm 3.8 triệu bản ghi. Sau lọc bản ghi không hợp lệ và loại bỏ ngoại lệ bằng z-score, còn 3.25 triệu bản ghi dùng huấn luyện. Việc làm sạch giúp giảm nhiễu, chuẩn hóa phân phối đặc trưng và cải thiện khả năng hội tụ.

Bảng 3.1 Thống kê mô tả bộ dữ liệu ban đầu sau lọc cơ bản

Metric	PULocationID	DOLocationID	passenger_count	trip_distance	pickup_day	pickup_hour	time_taken
count	3,359,242	3,359,242	3,359,242	3,359,242	3,359,242	3,359,242	3,359,242
mean	165.4313	164.8266	1.3213	3.5718	4.1124	14.4921	1106.1202
std	63.1270	69.2999	0.7587	88.1959	1.8840	5.7499	2058.3556
min	1	1	1	0.01	1	0	1.0
max	265	265	8	160,244.91	7	23	585,424.0

Bảng 3.2 Thống kê mô tả bộ dữ liệu sau khi loại bỏ ngoại lệ bằng z-score

Metric	PULocationID	DOLocationID	passenger_count	trip_distance	pickup_day	pickup_hour	time_taken
count	3,251,980	3,251,980	3,251,980	3,251,980	3,251,980	3,251,980	3,251,980
mean	165.5400	164.9288	1.2189	3.4797	4.1084	14.4863	1058.9179
std	63.0748	69.2589	0.4885	4.6164	1.8776	5.7418	848.2776
min	1	1	1	0.01	1	0	1.0
max	265	265	3	208.0	7	23	7,280.0

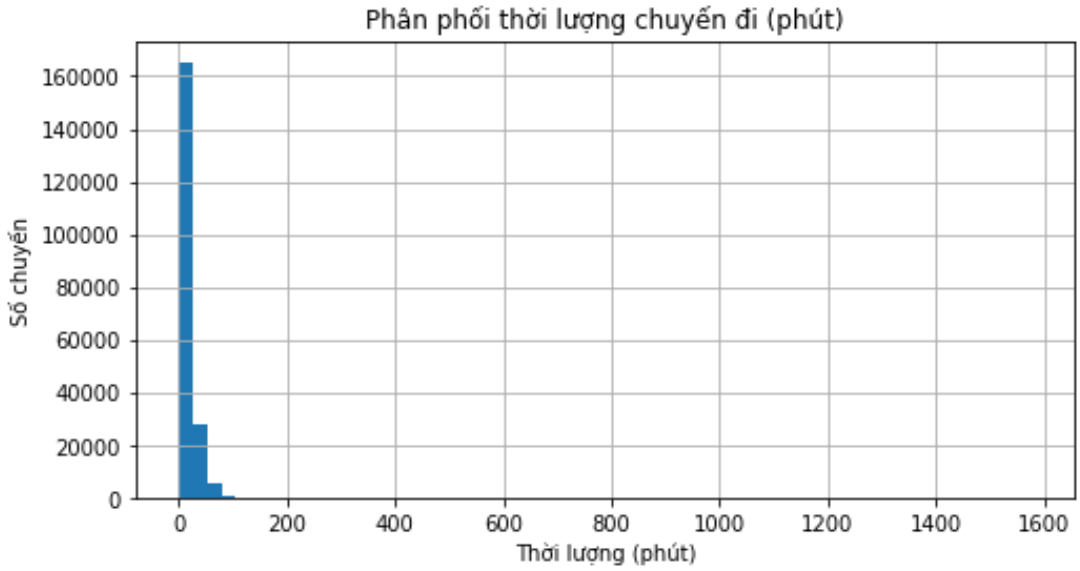
3.2.3 Biến đổi và tạo đặc trưng

Đặc trưng được mở rộng theo bốn nhóm: không gian (Taxi Zone Lookup), thời gian (chu kỳ), phân cụm vị trí (KMeans) và đặc trưng ngữ cảnh (tốc độ, thời lượng, phí tắc nghẽn). Bộ

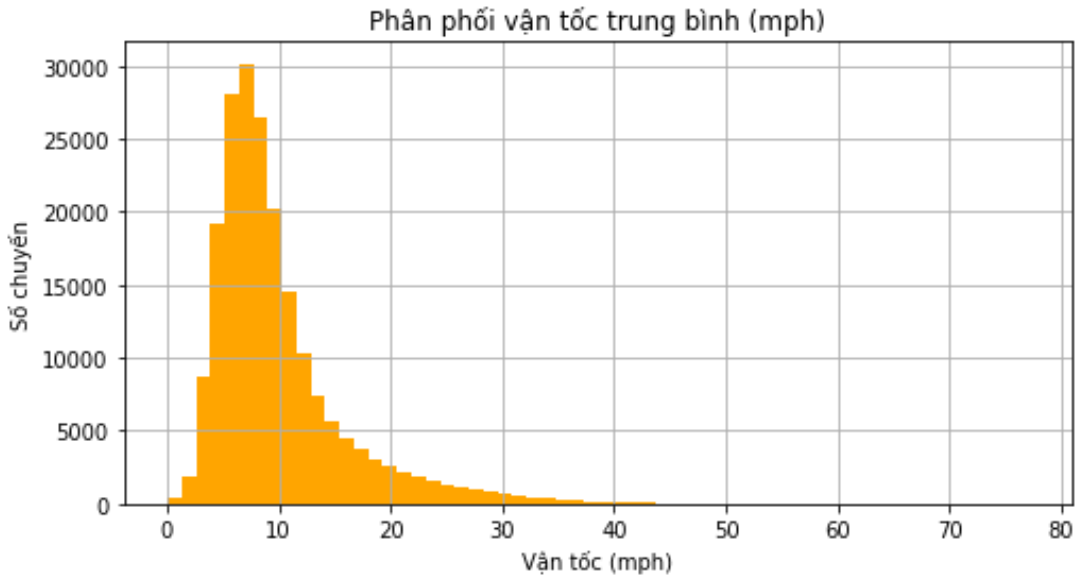
dữ liệu cuối cùng có hơn 30 đặc trưng, giúp mô hình học tốt hơn quan hệ phi tuyến.

Bảng 3.3 Danh sách các đặc trưng mới được tạo ra

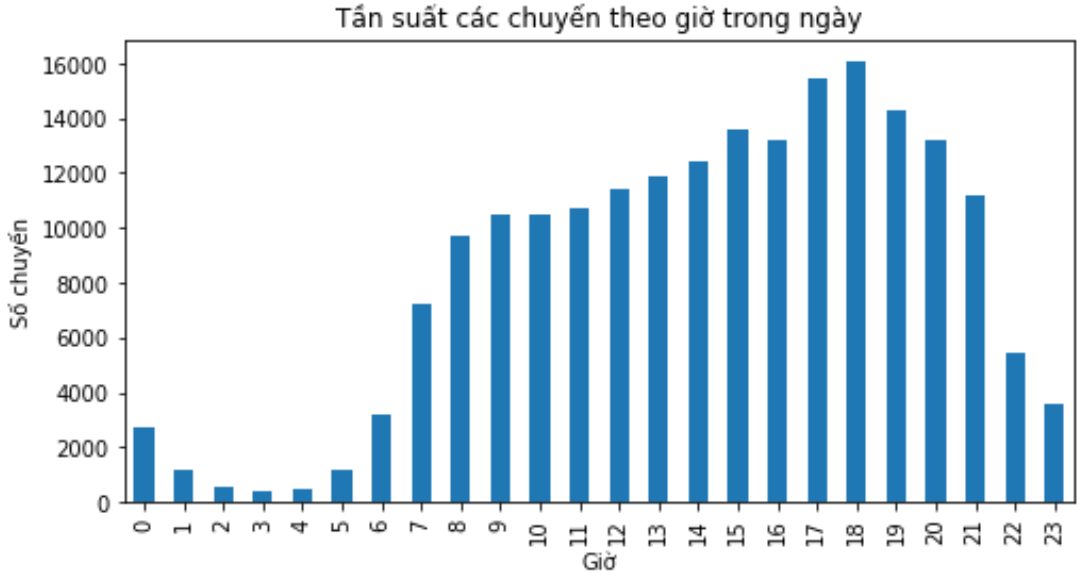
pickup_date	Thời gian	Ngày diễn ra chuyến đi (YYYY-MM-DD).
pickup_hour	Thời gian	Giờ đón khách (0–23).
pickup_minute	Thời gian	Phút trong giờ đón khách.
pickup_dow	Thời gian	Thứ trong tuần (1–7).
is_weekend	Thời gian	1 nếu là thứ 7 hoặc CN.
hour_sin, hour_cos	Chu kỳ	Mã hoá giờ trong ngày.
dow_sin, dow_cos	Chu kỳ	Mã hoá thứ trong tuần.
pickup_cluster	Không gian	KMeans trên PULocationID.
dropoff_cluster	Không gian	KMeans trên DOLocationID.
PU_Borough . . .	Lookup	Thông tin vùng đón.
DO_Borough . . .	Lookup	Thông tin vùng trả.
borough_pair	Không gian	Kết hợp Borough PU/DO.
OD_pair	Không gian	Ghép ID PU–DO.
is_manhattan_pu . . .	Nhị phân	Cờ khu vực Manhattan.
trip_duration_sec . . .	Ngữ cảnh	Thời lượng / vận tốc.



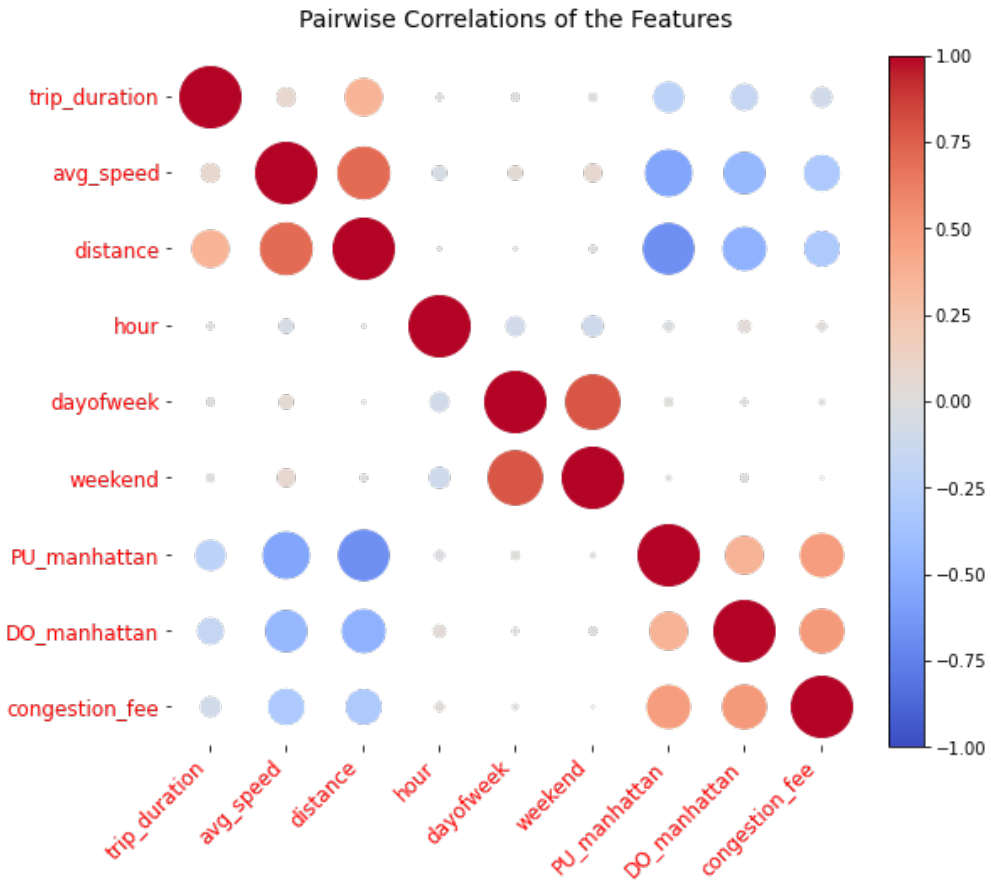
Hình 3.1 Phân phối thời lượng chuyến đi (đơn vị: phút)



Hình 3.2 Phân phối vận tốc trung bình của các chuyến taxi (mph)



Hình 3.3 Tần suất các chuyến taxi theo giờ trong ngày



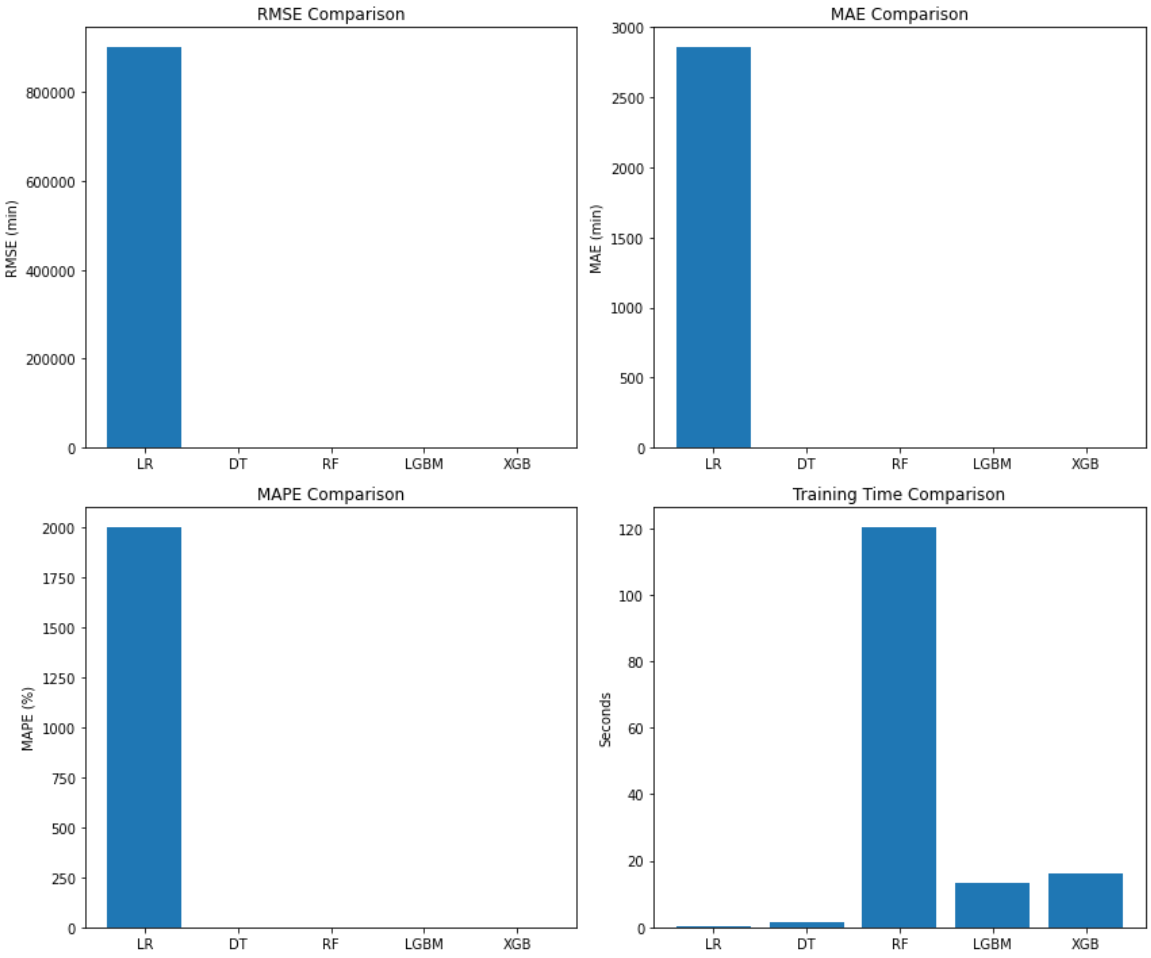
Hình 3.4 Ma trận tương quan giữa các đặc trưng chính

3.2.4 Huấn luyện mô hình

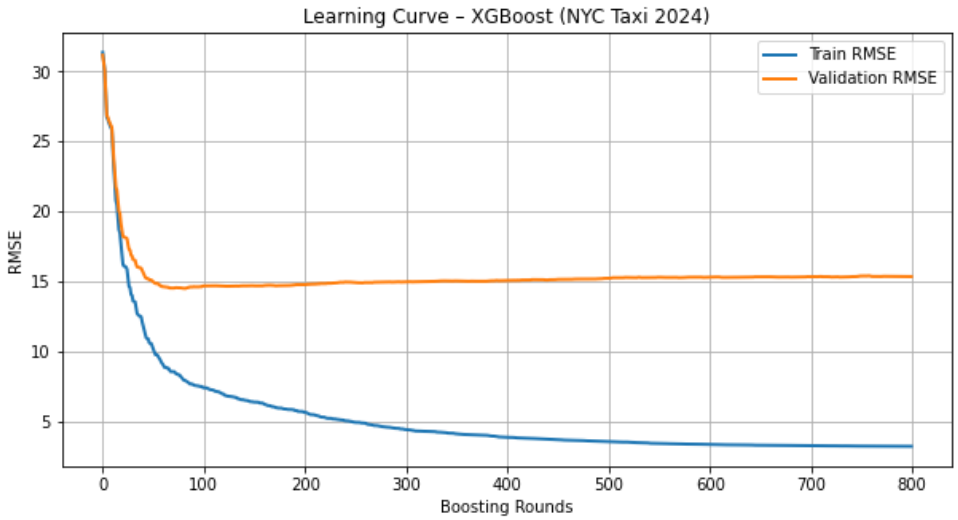
Dữ liệu được chia theo tỷ lệ 80/20. Mô hình XGBoost huấn luyện bằng SparkXGBRegressor với 800 cây, learning rate 0.08, max_depth=8. Việc song song hóa giúp huấn luyện trên 3.25M bản ghi trong thời gian hợp lý.

Bảng 3.4 So sánh hiệu năng giữa các mô hình trên tập kiểm thử

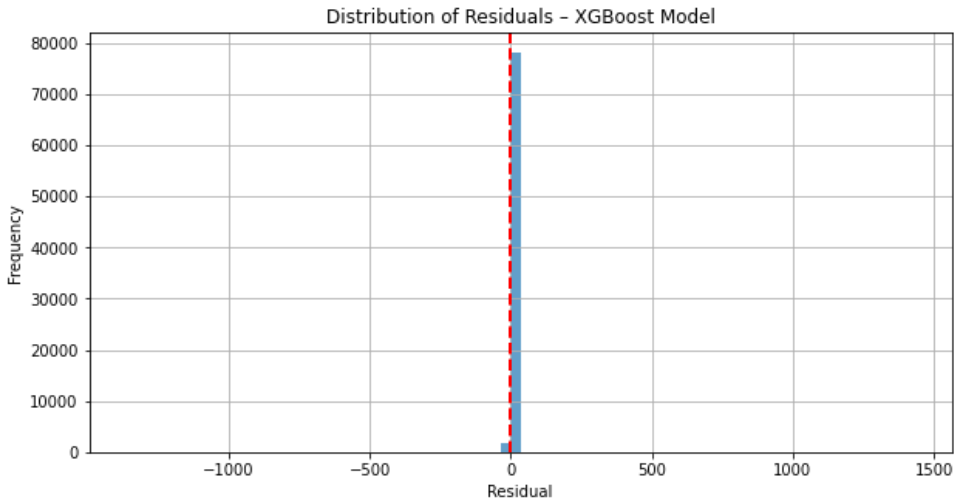
Mô hình	RMSE (min)	MAE (min)	MAPE (%)	Training time (s)
Linear Regression	900,481.15	2,857.56	2004.70	0.17
Decision Tree	17.66	0.69	1.57	1.43
Random Forest	23.33	0.72	0.86	120.38
LightGBM	18.16	0.61	1.65	13.52
XGBoost	**17.96**	**0.63**	**1.78**	**16.13**



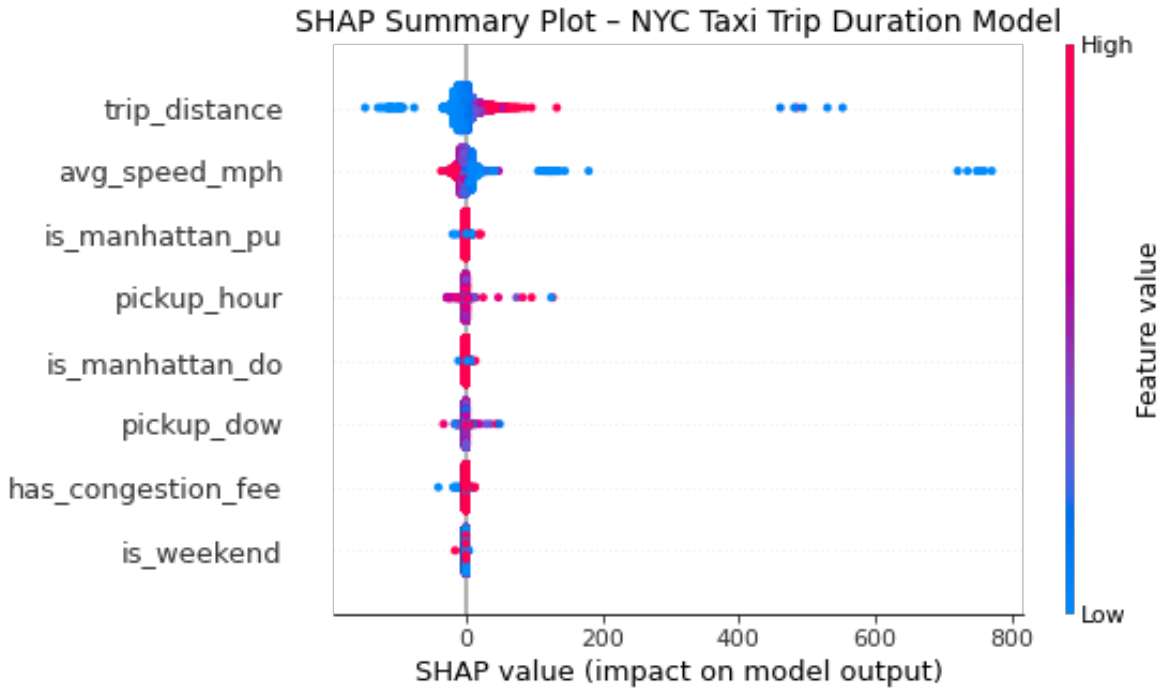
Hình 3.5 So sánh hiệu năng giữa các mô hình: RMSE, MAE, MAPE và thời gian huấn luyện



Hình 3.6 Đường hội tụ của mô hình XGBoost (Training vs Validation RMSE)



Hình 3.7 Phân phối sai số dự báo của mô hình XGBoost



Hình 3.8 Biểu đồ SHAP thể hiện mức độ đóng góp của các đặc trưng vào dự đoán

3.3 Kết luận chương

Chương này trình bày toàn bộ quy trình thực nghiệm: làm sạch dữ liệu, tạo đặc trưng, huấn luyện và đánh giá mô hình. XGBoost thể hiện vượt trội về độ chính xác và tốc độ, đồng thời SHAP giúp mô hình có khả năng diễn giải cao. Kết quả xác nhận XGBoost là mô hình phù hợp nhất cho bài toán ETA trên dữ liệu NYC Taxi 2024.

KẾT LUẬN

Đề án "Ứng dụng mô hình dự đoán chuỗi thời gian tối ưu chi phí vận tải" đã triển khai một quy trình nghiên cứu đầy đủ nhằm xây dựng và đánh giá các mô hình dự báo thời gian di chuyển (ETA) – một thành phần cốt lõi trong các hệ thống tối ưu hoá vận tải hiện đại. Thông qua ba chương lần lượt trình bày bối cảnh, phương pháp luận và kết quả thực nghiệm, đề án đã hoàn thành mục tiêu đặt ra: phát triển một mô hình dự báo thời gian di chuyển chính xác, ổn định và có khả năng ứng dụng thực tế trong việc giảm chi phí vận tải thông qua lập lịch phương tiện tối ưu.

Kết quả thực nghiệm cho thấy mô hình XGBoost đạt hiệu năng vượt trội nhất so với các mô hình đối chứng (Linear Regression, Decision Tree, Random Forest, LightGBM), đặc biệt trên các chỉ số quan trọng trong bài toán chuỗi thời gian như RMSE, MAE và MAPE. Nhờ khả năng học phi tuyến mạnh, tốc độ huấn luyện cao và tính ổn định khi xử lý dữ liệu nhiễu, XGBoost đáp ứng tốt yêu cầu của bài toán tối ưu chi phí vận tải – nơi mà chất lượng dự báo ETA tác động trực tiếp đến hiệu quả điều phối phương tiện, thời gian phục vụ và mức tiêu hao tài nguyên.

Đề án cũng đóng góp một quy trình xử lý dữ liệu toàn diện, bao gồm làm sạch, tích hợp dữ liệu vùng, tạo đặc trưng không gian–thời gian và bổ sung đặc trưng tắc nghẽn. Phân tích SHAP cho thấy các đặc trưng như *trip_distance*, *avg_speed_mph* và *pickup_hour* đóng vai trò quan trọng nhất trong dự báo thời gian di chuyển. Điều này giúp lý giải hoạt động của mô hình, đồng thời cung cấp nền tảng cho việc ra quyết định trong các bài toán vận tải như tối ưu tuyến đường, điều phối xe theo thời điểm và phân bổ nguồn lực theo khu vực.

Mặc dù đạt được nhiều kết quả tích cực, đề án vẫn tồn tại một số hạn chế. Bộ dữ liệu NYC Taxi 2024 không bao gồm thông tin GPS chi tiết, khiến việc mô hình hóa quỹ đạo phương tiện hoặc các mô hình học sâu theo chuỗi thời gian dài như ST-NN hoặc DeepTTE chưa thể triển khai đầy đủ. Ngoài ra, các yếu tố ngoại sinh như thời tiết, sự kiện đặc biệt hay tai nạn giao thông chưa được tích hợp vào hệ thống đặc trưng, trong khi đây là các yếu tố ảnh hưởng trực tiếp đến chi phí vận tải.

Dựa trên những hạn chế này, đề án đề xuất các hướng phát triển tiếp theo:

- Tích hợp các đặc trưng chính sách mới, đặc biệt là biến `cbd_congestion_fee` trong bộ dữ liệu TLC từ năm 2025 nhằm mô hình hóa chi phí tắc nghẽn trung tâm.
- Xây dựng mô hình dự báo chuyên biệt cho các điều kiện giao thông cực đoan để cải thiện độ ổn định và giảm rủi ro trong vận tải.
- Kết hợp XGBoost với các mô hình học sâu nâng cao như Graph Neural Networks (GNN) để mô hình hóa luồng giao thông và cấu trúc mạng lưới đô thị.
- Mở rộng nghiên cứu với dữ liệu GPS chi tiết hoặc dữ liệu cảm biến thời gian thực, từ đó

tối ưu thêm các thành phần chi phí như mức tiêu hao nhiên liệu, phân bổ đội xe, tối ưu tuyến đường động.

Tổng kết lại, đề án đã minh chứng rằng mô hình XGBoost là lựa chọn hiệu quả và cân bằng nhất cho bài toán dự đoán chuỗi thời gian trong tối ưu chi phí vận tải. Mô hình kết hợp tốt giữa độ chính xác, tốc độ huấn luyện và khả năng diễn giải, mở ra nhiều tiềm năng ứng dụng trong hệ thống vận tải thông minh và các dịch vụ logistics đô thị.