UIT Trường Đại học Công nghệ Thông tin | Khoa Khoa học và Kỹ thuật Thông tin

PHÂN TÍCH VÀ XÂY DỰNG MÔ HÌNH DỰ ĐOÁN THỜI GIAN CHUYẾN ĐI CỦA TAXI

MÔN HỌC: THIẾT KẾ VÀ PHÂN TÍCH THỰC NGHIỆM

GVHD: TS. Đỗ Trọng Hợp

Nhóm sinh viên thực hiên:

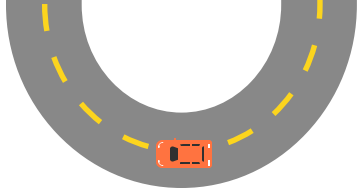
Trần Đại Hiển - *22520426*

Hồ Ngọc Mai - *22520839*

Võ Hoàng Thảo Phương - *22521171*

Lê Ngọc Thiên Phúc - *22521117*

DS304.O21



Nội dung đề tài

Giới thiệu

Bộ dữ liệu

Tiền xử lý dữ liệu

Thực nghiệm & mô hình

Kết quả thực nghiệm

Phân tích & hướng phát triển

Giới thiệu

Động lực

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.

Mục tiêu

- Nghiên cứu và đánh giá các yếu tố ảnh hưởng đến thời gian di chuyển của taxi.
- Xây dựng mô hình dự đoán thời gian chạy của taxi dựa trên các yếu tố được xác định.

Tổng quan đề tài

Bài toán: Phân tích và xây dựng mô hình dự đoán thời gian chuyến đi của taxi tại Chicago.
Input: Dữ liệu chuyến đi taxi tại Chicago.
Output: Thời gian chuyến đi của taxi tại Chicago

Kiểm định sự tác động của các yếu tố ảnh hưởng đến thời gian di chuyển của taxi bằng cách thêm các yếu tố như yếu tố thời tiết và yếu tố tắc nghẽn để cải thiện hiệu suất mô hình.

Bộ dữ liệu

Bộ dữ liệu **chuyến đi taxi tại Chicago** ban đầu, được thu thập từ trang **data.cityofchicago.org**, bao gồm **3,783,730 dòng** và **23 cột thuộc tính** thể hiện thông tin các chuyến taxi ở Chicago năm 2023.

	Trip ID	Taxi ID	Trip Start Timestamp	Trip End Timestamp	Trip Seconds	Trip Miles	Pickup Census Tract	Dropoff Census Tract	Pickup Community Area	Dropoff Community Area	...	Extras	Trip Total	Payment Type	Company	Pickup Centroid Latitude	Pickup Centroid Longitude	Pickup Centroid Location	Dropoff Centroid Latitude	Dropoff Centroid Longitude	Dropoff Centroid Location
0	0fca59218b11688279d795c03c4d16f851f13fa0	e2c349c7cbb608d552aa0b5814031943f13641ef9e50d8...	01/01/2023 12:00:00 AM	01/01/2023 12:15:00 AM	1037.0	4.82	NaN	NaN	6.0	32.0	...	0.0	19.50	Credit Card	Taxicab Insurance Agency Llc	41.944227	-87.655998	POINT (-87.6559981815 41.9442266014)	41.878866	-87.625192	POINT (-87.6251921424 41.8788655841)
1	1e539d6e7501164c6b76b761c3152c235e206d59	4ab7a7510c1ebcc9b2e3eaa7bdd6508dbea34da7986aca...	01/01/2023 12:00:00 AM	01/01/2023 12:15:00 AM	1341.0	16.63	NaN	NaN	76.0	8.0	...	6.0	53.00	Credit Card	Sun Taxi	41.980264	-87.913625	POINT (-87.913624596 41.9802643146)	41.899602	-87.633308	POINT (-87.6333080367 41.899602111)
2	2b3c5200439d51f626b60380809bbbcca766a85b	8c76eb82f069c0731a0049cb78898f02cc5ac6990244c9...	01/01/2023 12:00:00 AM	01/01/2023 12:15:00 AM	844.0	3.84	NaN	NaN	24.0	8.0	...	0.0	20.17	Mobile	Sun Taxi	41.901207	-87.676356	POINT (-87.6763559892 41.9012069941)	41.899602	-87.633308	POINT (-87.6333080367 41.899602111)
3	45b2ea39cfff64d61a46ef016e16f8ee74e9ed23	a688de71e9eb70603ba839dc7faf949968ae3e971e0575...	01/01/2023 12:00:00 AM	01/01/2023 12:00:00 AM	361.0	0.63	NaN	NaN	32.0	32.0	...	1.0	6.50	Cash	5 Star Taxi	41.878866	-87.625192	POINT (-87.6251921424 41.8788655841)	41.878866	-87.625192	POINT (-87.6251921424 41.8788655841)
4	464df6aaaf97ca8745985c2a5b2e481067a2bfb6	8b1a88e5a09cfd55ca72d267f00f56fa50a42aa322bdfe...	01/01/2023 12:00:00 AM	01/01/2023 12:15:00 AM	704.0	0.99	NaN	NaN	14.0	14.0	...	0.0	7.75	Cash	Flash Cab	41.968069	-87.721559	POINT (-87.7215590627 41.968069)	41.968069	-87.721559	POINT (-87.7215590627 41.968069)

Bộ dữ liệu **thời tiết Chicago,** được thu thập từ **kaggle.com,** bao gồm **24,108 dòng** và **10 cột thuộc tính** chứa thông tin thời tiết ở Chicago.

	YEAR	MO	DY	HR	TEMP	PRCP	HMDT	WND_SPD	ATM_PRESS	REF
0	2021	3	31	18	2.87	0.00	59.62	7.72	100.30	202103
1	2021	3	31	19	2.68	0.00	62.12	7.64	100.38	202103
2	2021	3	31	20	2.34	0.00	66.19	7.88	100.44	202103
3	2021	3	31	21	1.88	0.00	69.12	8.09	100.48	202103
4	2021	3	31	22	1.54	0.00	67.50	8.28	100.52	202103

Tiền xử lý dữ liệu

Làm sạch dữ liệu

- Xóa các dòng dữ liệu bị thiếu
- Loại bỏ dữ liệu nhiễu trên thuộc tính Trip Seconds
- Xóa bỏ các chuyến đi, đến từ các vùng ngoài Chicago
- Xóa bỏ các chuyến có điểm đón, trả khách giống nhau
- Tính khoảng cách pick-drop theo công thức Manhattan
- Thêm các đặc trưng về thời tiết từ bộ dữ liệu thời tiết
- Tạo thêm cột is_rush_hour đánh dấu cho giờ cao điểm

Raw Data

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

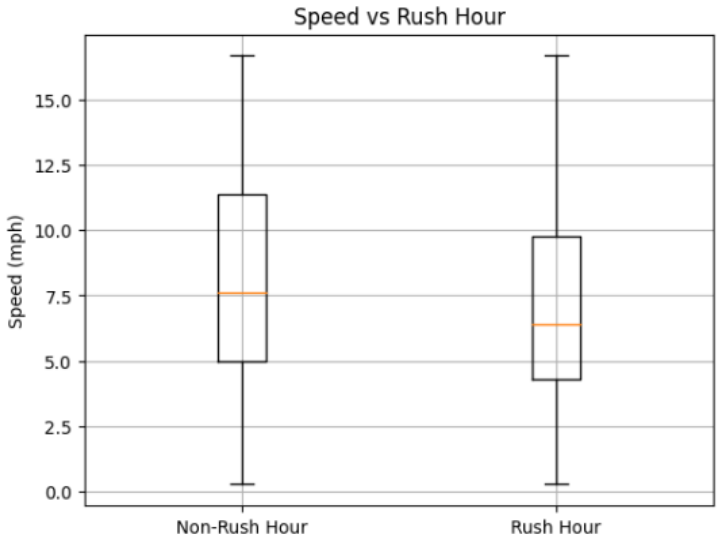
Cleaned Data

RangeIndex: 2253047 entries, 0 to 2253046
Data columns (total 15 columns):

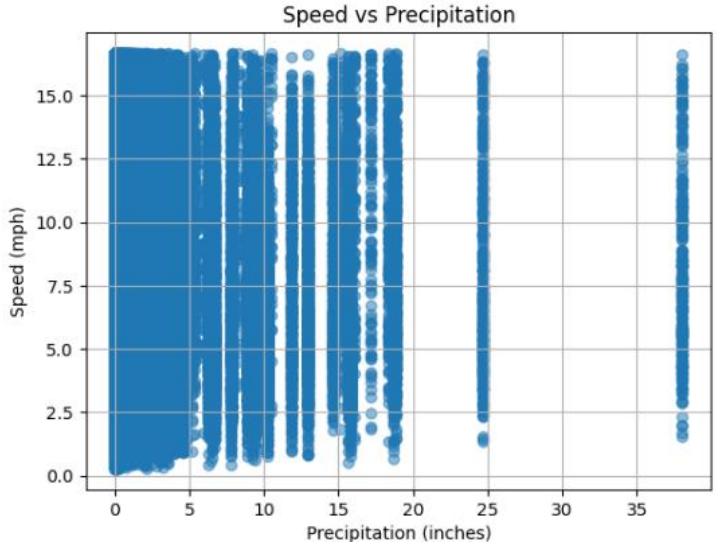
#	Column	Dtype
0	Trip Seconds	float64
1	Company	int64
2	Pickup Community Area	float64
3	Dropoff Community Area	float64
4	month	int64
5	hour	int64
6	day_of_year	int64
7	day_of_week	int64
8	day	int64
9	distance	float64
10	tempearture	float64
11	precipitation	int64
12	humidity	int64
13	wind_speed	int64
14	congestion	int64

Tiền xử lý dữ liệu

Khai phá dữ liệu



Ảnh hưởng của tắc nghẽn giao thông trong giờ cao điểm đến tốc độ xe



Ảnh hưởng của thời tiết đến tốc độ xe

Thực nghiệm & mô hình

Models:
- Linear Regression
- Ridge Regression

Data:
- Original
- Original + weather features
- Original + congestion features
- Original + weather features + congestion features

Độ đo:
- Root Mean Squared Logarithmic Error (RMSLE)

Cross-validation: 10-fold

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

```
original_dataset: Mean RMSLE = 0.3729
original_dataset+weather_features: Mean RMSLE = 0.3713
original_dataset+congestion_features: Mean RMSLE = 0.3695
original_dataset+weather+congestion_features: Mean RMSLE = 0.3674
```

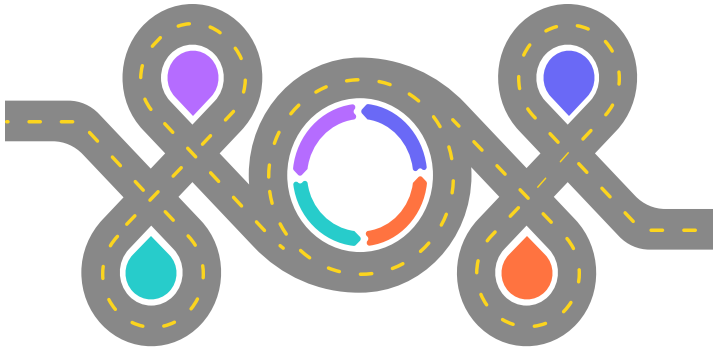
Phân tích

ANOVA

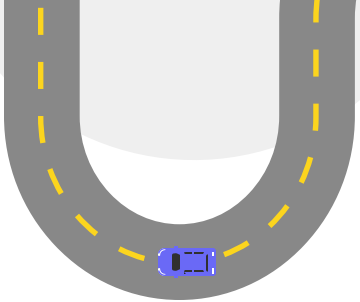
Lấy mức ý nghĩa: 0.05

```
              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...
```

- H0: Không có sự khác biệt về kết quả của các nhóm
- H1: Có sự khác biệt về kết quả của các nhóm



-> Post-hoc analysis



Post-hoc analysis

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

```
original_dataset: Mean RMSLE = 0.3729
original_dataset+weather_features: Mean RMSLE = 0.3713
original_dataset+congestion_features: Mean RMSLE = 0.3695
original_dataset+weather+congestion_features: Mean RMSLE = 0.3674
```

Kết luận và hướng phát triển

Models

Feature selection

New features

CẢM ƠN THẦY & CÁC BẠN ĐÃ LẮNG NGHE!